Stability and Non-Stationary Characteristics of Queues

A Thesis
Presented to
The Academic Faculty

by

Brian H. Fralix

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Industrial and Systems Engineering
Georgia Institute of Technology
May 2007

Stability and Non-Stationary Characteristics of Queues

Approved by:

Dr. Richard F. Serfozo, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Christian Houdré
School of Mathematics
Georgia Institute of Technology

Dr. Hayriye Ayhan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Alexander Shapiro
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Robert D. Foley
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: December 18, 2006

To my family

ACKNOWLEDGEMENTS

I would like to begin by thanking the members of my thesis committee: Prof. Hayriye Ayhan, Prof. Robert Foley, Prof. Christian Houdré, and Prof. Alex Shapiro. Their comments, criticisms and suggestions have proven to be beneficial for my research, and I have greatly benefitted from attending courses they have taught.

I should also mention my colleagues. I won't begin to name everyone, for fear of exclusion, but I would like to thank my fellow grad students for providing an environment that made my grad school experience much more tolerable.

Last, but certainly not least, I would like to thank my advisor, Prof. Richard Serfozo. It is impossible for me to express in words my full appreciation, but I would like to give thanks for all of the advice and encouragement he has given throughout my Ph.D. studies, and for putting up with my "hard-headed" ways. All of the students he has advised in the past bring attention to both his pleasant demeanor and his seemingly infinite patience with his students. I would like to supplement this by bringing the reader's attention to the fact that he was willing to advise me through the first 1.5 years of his retirement; he could have easily said no at the earlier stages of my graduate studies. It is for these things, and many others, that I am truly grateful.

TABLE OF CONTENTS

SUMMARY

We provide contributions to two classical areas of queueing. The first part of this thesis focuses on finding new conditions for a Markov chain on a general state space to be Harris recurrent, positive Harris recurrent or geometrically ergodic. Most of our results show that establishing each property listed above is equivalent to finding a good enough feasible solution to a particular optimal stopping problem, and they provide a more complete understanding of the role Foster's criterion plays in the theory of Markov chains.

The second and third parts of the thesis involve analyzing queues from a transient, or time-dependent perspective. In part two, we are interested in looking at a queueing system from the perspective of a customer that arrives at a fixed time t. Doing this requires us to use tools from Palm theory. From an intuitive standpoint, Palm probabilities provide us with a way of computing probabilities of events, while conditioning on sets of measure zero. Many studies exist in the literature that deal with Palm probabilities for stationary systems, but very few treat the non-stationary case. As an application of our main results, we show that many classical results from queueing (in particular *ASTA* and Little's law) can be generalized to a time-dependent setting.

In part three, we establish a continuity result for what we refer to as jump processes. From a queueing perspective, we basically show that if the primitives

and the initial conditions of a sequence of queueing processes converge weakly, then the corresponding queue-length processes converge weakly as well in some sense. Here the notion of convergence used depends on properties of the limiting process, therefore our results generalize classical continuity results that exist in the literature. The way our results can be used to approximate queueing systems is analogous to the way phase-type random variables can be used to approximate other types of random variables.

CHAPTER I

INTRODUCTION

Queues are prevalent within our society. Many examples of queues can be found throughout the various manufacturing and production systems that exist today. Even people that don't work in a traditional industrial setting observe queueing systems in action: they participate in them while visiting a doctor's office, while waiting at a traffic light, or while shopping at a supermarket. Therefore, it is important that we learn as much as we can about queues, in order to improve our quality of life.

When designing a queue, it is important to understand whether or not it will be able to handle all of the demands placed upon it. This is a question of stability, which will be the topic of discussion in the first part of this thesis. The second and final parts of the thesis will involve trying to understand how a queue behave in a "local" sense, in other words, how it behaves at a fixed time t.

To be more specific, the first contribution of this thesis involves finding new conditions for a Markov chain on a general state space to be Harris recurrent, positive Harris recurrent or geometrically ergodic. Our results are stated in terms of drift conditions that are similar to the standard drift conditions used to verify these criteria, but our conditions involve random steps. In particular, our results imply that showing a Markov chain is stable is equivalent to finding a "good enough" feasible solution to an optimal stopping problem. Filonov gives sufficient conditions for a

countable-state Markov chain to be positive recurrent, and we show how to use his ideas to extend results to the general state-space setting. While the results themselves do not explicitly refer to a queue, it may be possible to use them as a first step towards finding stability conditions for some queues in the literature, since it is well-known that Markov processes can be used to model many interesting types of queueing systems.

The second part of the thesis involves showing how non-stationary Palm probabilities can be used to analyze transient properties of queues. We first show that, even for non-stationary point processes, a point process has a $\mathcal{F}_t$-intensity if and only if (almost all of) the induced Palm probabilities P_t are absolutely continuous with respect to the underlying measure P on the σ-field $\mathcal{F}_{t-}$. This result is used to provide necessary and sufficient conditions for a transient version of *ASTA* (Arrivals See Time Averages) to hold at a time t. What's interesting about this result is that the conditions are essentially the same, structurally, as the conditions given by Melamed and Whitt in the stationary setting. Next, we present some local and long-run limit results that pertain to our Palm probabilities, and we use these results to show that a classical heuristic argument used to prove *PASTA* can actually be transformed into a rigorous argument. After proving more results pertaining to Palm probabilities within the context of Markov jump processes, we conclude this discussion by showing how Palm theory can be used to provide quick proofs of the transient versions of Little's law that were derived by Bertsimas and Mortzinou. Moreover, the use of Palm probabilities also allows us to relate *any* moment of the queue-length of time t to the waiting times experienced by customers that arrive before time t. The reader should

realize that our use of Palm probabilities still only allow us to come to conclusions that are somewhat qualitative. In other words, they only lead to very broad (but still useful) relationships between various distributions within our processes. Actually computing some of the expressions given here, such as the expected waiting time of a customer that arrives at a time t, would involve a heavy use of numerical methods, and this alone could serve as the subject of another study.

The final part of the thesis continues the study of transient properties of queueing systems. In this section, we provide a continuity result for queue-length processes. Our results can be used to prove that, for a sequence $\{Q_n(t) : t \geq 0\}$ of $GI/GI/1$ queues with interarrival distribution A_n and service distribution S_n, $A_n \Rightarrow A$ and $S_n \Rightarrow S$ implies that $Q_n \Rightarrow Q$ with respect to the Skorohod topology, so long as $\{Q(t) : t \geq 0\}$ is the queue-length process of a $GI/GI/1$ queue with continuous interarrival and service time distributions A and B, respectively. For general A and B, the Skorohod topology cannot be used, due to the fact that arrivals and services may occur simultaneously. To account for this, we define a new notion of convergence that allows for such behavior, and we prove another continuity result that allows us to approximate such queues under this weaker (but necessary) notion of convergence. These results approximate queues in a way that's analogous to approximating a random variable with a phase-type random variable.

CHAPTER II

FOSTER-TYPE CRITERIA

2.1 Introduction

The classical Foster criterion is well-known for verifying whether an irreducible Markov chain on a countable state space is positive recurrent. Intuitively, this criterion assumes that the chain tends to drift (in unit steps) towards some small subset of the state space, and the chain doesn't wander too far when it makes a one-step transition out of this set.

This chapter addresses the following issue. Can one find analogous drift criteria for Markov chains on general state spaces that are based on steps that may be larger than one or on random steps? Specifically, are there drift criteria that are sufficient for Harris recurrence, positive Harris recurrence, or geometric ergodicity?

The first study that addresses such issues was by Filonov [13] (also see [25]). He gives a sufficient drift condition for a Markov chain on a countable space to be ergodic for steps that are stopping times. Meyn and Tweedie [22] obtained similar results for Markov chains on arbitrary state spaces for deterministic steps. They also derive a sufficient condition involving steps that are conditionally independent of the Markov chain (given a fixed initial state), with tail probabilities that satisfy a certain property. Because of this property, their random state-dependent drift criterion is not a generalization of their deterministic version.

In this chapter we present new Foster-type drift conditions involving steps that are stopping times with respect to a filtration that preserves the Markovian property of the Markov chain under study. Our first result (Theorem 3) provides a sufficient condition for a Markov chain on a general state space to be Harris recurrent. It is based on a drift condition for steps that are stopping times with respect to a suitable filtration. Theorems 2.1(i) and 2.2(i) in [22] are special cases. Our next result (Theorem 5) is a sufficient condition for positive Harris recurrence under drift conditions for steps that are integrable stopping times. Theorem 5 generalizes many known results, including Filonov's result [13] (for a countable state space), Theorem 2.1(ii) in [22] (for deterministic steps) and Theorem 1 in [14]. However, Theorem 5 and Theorem 2.1(ii) are not comparable for some cases; see the example in section 2 of [22] for further insight on this. Our next result (Theorem 6) gives a sufficient condition for a Markov chain to be geometrically ergodic. The special case of this result for deterministic steps is Theorem 2.1(iii) of [22]. We conclude by giving another state-dependent drift condition for geometric ergodicity that's similar to Theorem 1 in [14]. This result also generalizes Theorem 2.1(iii) of [22], but in a different way.

An important feature of our Theorems 5, 6, and 7 for establishing positive Harris recurrence and geometric ergodicity is that they do not require the Markov chain to be ψ-irreducibile. Costa and Dufour [9] also showed the ψ-irreducibility assumption is not needed for the one-step drift condition for positive Harris recurrence. Another sublety in their drift condition is that it uses extended real-valued test functions — our conditions, and those in [22], simply use real-valued test functions.

2.2 Preliminaries

We will study the type of Markov chain discussed in [21]. Let $X := \{X_n\}_{n=0}^{\infty}$ be a Markov chain with respect to a filtration $\mathcal{F} := \{\mathcal{F}_n\}_{n=0}^{\infty}$ on an arbitrary state space $\mathbb{E}$ equipped with a countably generated σ-field $\mathcal{E}$ (e.g., $\mathbb{E}$ could be a Polish space equipped with its Borel σ-field $\mathcal{E}$). The underlying probability space for the process is $(\Omega, \mathcal{A}, P)$.

We will frequently use stopping times of the form $\tau_A := \inf\{n \geq 1 : X_n \in A\}$, $A \in \mathcal{E}$. A set $C \in \mathcal{E}$ is *petite* if there exist a nontrivial measure μ on $\mathcal{E}$ and a probability measure α on the nonnegative integers such that

$$K_\alpha(x, B) := \sum_{n=0}^{\infty} \alpha(n) P_x(X_n \in B) \geq \mu(B), \qquad x \in C,\ B \in \mathcal{E}.$$

The Markov chain X is *ψ-irreducible* if there exists a nontrivial measure ψ on $\mathcal{E}$ such that $P_x(\tau_A < \infty) > 0$, $x \in \mathbb{E}$, for any $A \in \mathcal{E}$ with $\psi(A) > 0$. The Markov chain X is *Harris recurrent* if it is ψ-irreducible and $P_x(\tau_A < \infty) = 1$, $x \in A$, for any $A \in \mathcal{E}$ with $\psi(A) > 0$.

A measure π is *invariant* for X if $\pi(A) = \int_E P_x(X_1 \in A)\pi(dx)$, $A \in \mathcal{E}$. A Harris recurrent Markov chain has a unique invariant measure (up to constant multiple), and the chain is *positive Harris recurrent* if the measure is finite.

The Markov chain X is *geometrically ergodic* if there exists a function $M : E \to R_+$ and $\rho < 1$ such that

$$\sup_{A \in \mathcal{E}} |P_x(X_n \in A) - \pi(A)| \leq M(x)\rho^n, \qquad n \geq 1.$$

We will use the following theorems from [22].

Theorem 1 *Suppose the Markov chain X is ψ-irreducible.*

(i) ([22], Theorem 3.1(i)) X is Harris recurrent if there is a petite set $C \in \mathcal{E}$ such that $P_x(\tau_C < \infty) = 1$, for all $x \in \mathbb{E}$.

(ii) ([22], Theorem 3.2(i)) X is positive Harris recurrent if and only if there is a petite set $C \in \mathcal{E}$ such that $P_x(\tau_C < \infty) = 1$, for all x, and $\sup_{x \in C} E_x(\tau_C) < \infty$.

(iii) ([22], Theorem 3.3(ii)) X is geometrically ergodic if it is aperiodic and there exist a petite set $C \in \mathcal{E}$ and $\kappa > 1$ such that $E_x(\kappa^{\tau_C}) < \infty$, for all x, and $\sup_{x \in C} E_x(\kappa^{\tau_C}) < \infty$.

Theorem 2 *([22], Theorem 3.1(ii)) Suppose X is ψ-irreducible. There is a set $N \in \mathcal{E}$ such that N^c (the complement of N) is empty or it is absorbing, and X restricted to N^c is Harris recurrent and $\psi(N) = 0$. If X is not Harris recurrent, then N is nonempty and, for any petite set $C \subseteq N$ and $x \in N$,*

$$P_x(X_n \in N, n \geq 0) > 0, \qquad P_x(X_n \in C \ \ i.o.) = 0.$$

2.3 Results

For the following results, we assume the underlying probability space for the Markov chain X has the additional structure that there is a shift operator $\theta : \Omega \to \Omega$ such that $X_n(\omega) = X_0(\theta_n \omega)$, where $\theta_0 = I$ (the identity mapping on Ω) and $\theta_n = \theta \circ \theta_{n-1}$, $n \geq 1$. If $(\Omega, \mathcal{A})$ is the canonical probability space (i.e. the sequence space $\mathbb{E}^\infty$ equipped with the product σ-field), then θ is just the usual shift operator for sequences.

Associated with a stopping time τ, define random variables $0 = \tau_0 < \tau_1 < \ldots$ on $(\Omega, \mathcal{A})$ by

$$\tau_{n+1} = \tau_n + \tau \circ \theta_{\tau_n}, \qquad n \geq 0.$$

Note that each τ_n is an $\mathcal{F}$-stopping time, and by the strong Markov property (which holds because our time index is discrete), the process $\overline{X}_n := X_{\tau_n}$, $n \geq 0$, is also a Markov chain with respect to the filtration $\mathcal{F}^\tau := \{\mathcal{F}_{\tau_n}\}_{n=0}^\infty$. In addition to the first entrance time τ_A to the set A, we will use

$$\sigma_A := \inf\{n \geq 0 : X_n \in A\}, \qquad \overline{\sigma}_C := \inf\{n \geq 0 : \overline{X}_n \in C\}.$$

Our first result provides a sufficient condition for X to be Harris recurrent.

Theorem 3 *Suppose the Markov chain X is ψ-irreducible and there exist an $f : \mathbb{E} \to R_+$ that is unbounded off petite sets, a finite $\mathcal{F}$-stopping time $\tau \geq 1$, and a petite set C such that*

$$E_x[f(X_\tau)] \leq f(x), \qquad x \notin C. \tag{1}$$

Then X is Harris recurrent.

Proof Proceeding as in the proof of Theorem 2.1(i) in [22], define $U_n = f(\overline{X}_n)\mathbf{1}(\overline{\sigma}_C \geq n)$. Clearly $\overline{\sigma}_C$ is an $\mathcal{F}^\tau$-stopping time. This fact, along with our drift condition (1) gives

$$E[U_n|\mathcal{F}_{\tau_{n-1}}] = \mathbf{1}(\overline{\sigma}_C \geq n)E[f(\overline{X}_n)|\mathcal{F}_{\tau_{n-1}}] \leq U_{n-1}.$$

This implies that $U := \{U_n\}_{n=0}^\infty$ is a nonnegative supermartingale with respect to the filtration $\mathcal{F}^\tau$, so it converges a.s. to a finite limit. Notice that if $\overline{\sigma}_C < \infty$ a.s., then the limit must be zero a.s..

Consider the set N given in Theorem 2, and suppose that it is nonempty (otherwise we would be done). Notice that while on the set $\Omega_N = \{X_k \in N, k \geq 0\}$, we only have to consider the case when $\lim_{n\to\infty} f(X_n) = \infty$ a.s. on Ω_N, since f is unbounded

off petite sets and $P_x(X_n \in G \ \ i.o.) = 0$ for any petite set $G \subseteq N$, and $x \in N$, by Theorem 2 (if the limit inferior is finite with positive probability, the contrapositive of the second part of Theorem 2 allows us to conclude that the chain is Harris recurrent). Therefore, $\lim_{n\to\infty} f(\overline{X}_n) = \infty$ a.s. on Ω_N, which means that $\overline{\sigma}_C < \infty$ a.s. on Ω_N (notice the use of the supermartingale U). Since $\tau_n < \infty$ a.s. for each n, we also know that $\sigma_C < \infty$ a.s. on Ω_N.

Now assume that N^c is nonempty (if it is empty, then our proof is complete by Theorem 1(i)). From Theorem 2, we know there is a petite set $D \subset N^c$ such that $\psi(D) > 0$, since $\psi(N) = 0$. From Harris recurrence on N^c, it follows that $P_x(\sigma_D < \infty) = 1$, $x \in N^c$. However, for all paths not in Ω_N, it follows that $\sigma_D < \infty$ a.s. for any initial point y, because N^c is an absorbing set.

¿From Theorem 5.5.5. of [21], we know that $C \cup D$ is petite. Thus, for any $x \in \mathbb{E}$,

$$P_x(\sigma_{C\cup D} = \infty) \leq P_x(\sigma_C = \infty, \Omega_N) + P_x(\sigma_D = \infty, \Omega_N^c) = 0,$$

which completes the proof (use Theorem 1(i)). ∎

The following lemma will be used in the rest of our proofs.

Lemma 4 *If there is a petite set C such that $P_x(\tau_C < \infty) > 0$, for all $x \in \mathbb{E}$, then the Markov chain X is ψ-irreducible.*

Proof Since C is petite, there exist a probability measure $\alpha(\cdot)$ and a nontrivial measure ψ such that

$$K_\alpha(x, B) \geq \psi(B), \qquad x \in C, \ B \in \mathcal{E}.$$

Fix $y \in \mathbb{E}$. Since $P_y(\tau_C < \infty) > 0$, there exists an integer n_0 such that $P_y(X_{n_0} \in C) > 0$. Thus, for any set $B \in \mathcal{E}$ such that $\psi(B) > 0$,

$$\begin{aligned}
\sum_{m=n_0}^{\infty} P_y(X_m \in B) &= \int_{\mathbb{E}} \left(\sum_{m=0}^{\infty} P_x(X_m \in B) \right) P_y(X_{n_0} \in dx) \\
&\geq \int_C K_\alpha(x, B) P_y(X_{n_0} \in dx) \\
&\geq \psi(B) P_y(X_{n_0} \in C) > 0.
\end{aligned}$$

So by Proposition 4.2.1 in [21], X is ψ-irreducible. ∎

The following result is an extension of Theorem 1 in Foss and Konstantopoulos [14] to the random drift setting. Conditions (4)-(7) below are similar to (L1)-(L4) in [14], respectively.

Theorem 5 *X is positive Harris recurrent if there exist a petite set C, an $f : \mathbb{E} \to R_+$ bounded on C, constants A and α, a stopping time $\tau \geq 1$ a.s., and a $\mathcal{F}_\infty$-measurable random variable η that satisfy the following:*

$$E_x[f(X_\tau) + \eta] \leq f(x), \qquad x \notin C \tag{2}$$

$$\sup_{x \in C} E_x[f(X_\tau) + \eta] < \infty \tag{3}$$

$$\eta(\omega) \geq \alpha > -\infty, \qquad \omega \in \Omega \tag{4}$$

$$\eta(\omega) > 0, \qquad X_0(\omega) \notin C \tag{5}$$

$$\sup_{x \in C} E_x[\tau] < \infty \tag{6}$$

$$\tau(\omega) \leq A\eta(\omega), \qquad X_0(\omega) \notin C \tag{7}$$

This theorem is a generalization of many standard methods. When $\eta = \tau$, we have the general state-space analogue of Filonov's result. In particular, if $\eta = \tau = g(X_0)$ for some function g and f is bounded on C, we end up with Meyn and Tweedie's criterion found in [22]. Another special case of Theorem 5 is Dai's method [10], which involves the use of fluid limits; see [14] for details on how this method is equivalent to satisfying certain drift criteria.

Proof In addition to our sequence $\{\tau_n\}_{n=0}^{\infty}$, we introduce another sequence $\{\eta_n\}_{n=0}^{\infty}$, where $\eta_0 = 0$ and for $n \geq 0$, $\eta_{n+1} = \eta_n + \eta \circ \theta_{\tau_n}$. Note that, by induction,

$$\eta_n = \sum_{k=0}^{n-1} \eta \circ \theta_{\tau_k}. \tag{8}$$

Let $\nu := \inf\{n \geq 0 : \tau_n \geq \sigma_C\}$. This is an $\mathcal{F}^\tau$-stopping time since it is easy to see that $\{\nu \leq n\} \in \mathcal{F}_{\tau_n}$.

We will now show that $\{Y_{\nu\wedge n}\}_{n=0}^{\infty}$ is an $\mathcal{F}^\tau$-supermartingale, where $Y_n := f(\overline{X}_n) + E[\eta_n|\mathcal{F}_{\tau_n}]$. Notice that on the set $\{\nu > n\}$,

$$\begin{aligned} E[Y_{n+1}|\mathcal{F}_{\tau_n}] &= E[f(X_{\tau_n+\tau\circ\theta_{\tau_n}}) + \eta \circ \theta_{\tau_n}|\mathcal{F}_{\tau_n}] + E[\eta_n|\mathcal{F}_{\tau_n}] \\ &= E_{\overline{X}_n}[f(X_\tau) + \eta] + E[\eta_n|\mathcal{F}_{\tau_n}] \\ &\leq f(\overline{X}_n) + E[\eta_n|\mathcal{F}_{\tau_n}] = Y_n. \end{aligned}$$

Thus, $\{Y_{\nu\wedge n}\}_{n=0}^{\infty}$ is a nonnegative $\mathcal{F}^\tau$-supermartingale. Since f is nonnegative,

$$E_x[\eta_{\nu\wedge n}] = E_x[E_x[\eta_{\nu\wedge n}|\mathcal{F}_{\tau_{\nu\wedge n}}]] \leq E_x[Y_{\nu\wedge n}] \leq f(x). \tag{9}$$

Using (8),

$$\eta_{\nu\wedge n} = \sum_{k=0}^{n-1} \eta \circ \theta_{\tau_k} \mathbf{1}(k < \nu)$$

$$\begin{aligned} &\geq A^{-1}\sum_{k=0}^{n-1}\tau\circ\theta_{\tau_k}\mathbf{1}(k<\nu) \\ &= A^{-1}\tau_{\nu\wedge n}. \end{aligned}$$

The first inequality follows from property (7), since $\nu > k$ implies $X_0(\theta_{\tau_k}\omega) \notin C$. After taking limits, while using (9) and $\sigma_C \leq \tau_\nu$, we see that for $x \notin C$,

$$E_x[\tau_C] = E_x[\sigma_C] \leq Af(x). \tag{10}$$

For $x \in C$,

$$E_x[\tau_C] = E_x[\tau_C\mathbf{1}(\tau_C \leq \tau)] + E_x[\tau_C\mathbf{1}(\tau_C > \tau)]. \tag{11}$$

Now, by (10)

$$\begin{aligned} E_x[\tau_C\mathbf{1}(\tau_C>\tau)] &\leq E_x[\tau\mathbf{1}(\tau_C>\tau)] + E_x[\mathbf{1}(\tau_C>\tau)E_{\overline{X}_1}[\tau_C]] \\ &\leq E_x[\tau\mathbf{1}(\tau_C>\tau)] + E_x[\mathbf{1}(\tau_C>\tau)Af(\overline{X}_1)] \\ &\leq E_x[\tau\mathbf{1}(\tau_C>\tau)] + AE_x[f(\overline{X}_1)+\eta-\eta] \end{aligned}$$

Using this with (3), (6) and (11), we have

$$E_x[\tau_C] \leq E_x[\tau] + AE_x[f(\overline{X}_1)+\eta] - A\alpha < \infty.$$

Then X is positive Harris recurrent by Lemma 1 and Theorem 1(i).

∎

Our next result provides sufficient conditions for X to be geometrically ergodic.

Theorem 6 *The Markov chain X is geometrically ergodic if it is aperiodic and there exists an $f : \mathbb{E} \to [1, \infty)$, an $\mathcal{F}$-stopping time $\tau \geq 1$, a petite set C, and a constant $\kappa > 1$ such that*

$$\begin{aligned} E_x[\kappa^\tau f(X_\tau)] &\leq f(x), \qquad x \notin C \\ \sup_{x \in C} E_x[\kappa^\tau f(X_\tau)] &< \infty, \qquad x \in C. \end{aligned} \tag{12}$$

Proof This proof is similar to the proof of Theorem 5. Let $Y_n := \kappa^{\tau_n} f(X_{\tau_n})$. Then on the set $\{\nu > n\}$,

$$\begin{aligned} E[Y_{n+1}|\mathcal{F}_{\tau_n}] &= E[\kappa^{\tau_n + \tau \circ \theta_{\tau_n}} f(X_{\tau_n + \tau \circ \theta_{\tau_n}})|\mathcal{F}_{\tau_n}] \\ &= \kappa^{\tau_n} E_{X_{\tau_n}}[\kappa^\tau f(X_\tau)] \leq Y_n. \end{aligned}$$

This shows that $\{Y_{\nu \wedge n}\}_{n=0}^\infty$ is a nonnegative supermartingale with respect to $\mathcal{F}^\tau$. Next, observe that

$$E_x(\kappa^{\tau_{\nu \wedge n}}) \leq E_x(Y_{\nu \wedge n}) \leq E_x(Y_0) = f(x), \qquad x \in \mathbb{E}.$$

Letting $n \to \infty$ and applying the monotone convergence theorem, we have

$$E_x(\kappa^{\sigma_C}) \leq E_x(\kappa^{\tau_\nu}) \leq f(x), \qquad x \in \mathbb{E}.$$

Then, for $x \notin C$, $E_x[\kappa^{\tau_C}] = E_x[\kappa^{\sigma_C}] \leq f(x) < \infty$. And for $x \in C$,

$$\begin{aligned} E_x[\kappa^{\tau_C}] &\leq E_x[\kappa^\tau E_{X_\tau}[\kappa^{\sigma_C}]\mathbf{1}(\tau_C > \tau) + \kappa^\tau \mathbf{1}(\tau_C \leq \tau)] \\ &\leq E_x[\kappa^\tau f(X_\tau)]. \end{aligned}$$

Combining these observations yields $E_x[\kappa^{\tau_C}] < \infty$ for all $x \in \mathbb{E}$, so again Lemma 1 tells us that X is ψ-irreducible. Therefore, X is geometrically ergodic by Theorem

1(iii). ∎

It is also worth pointing out that a "supermartingale" approach helps to derive another drift condition for geometric ergodicity that's analogous to Theorem 1 in [14].

Theorem 7 *The Markov chain X is geometrically ergodic if it is aperiodic and there exist a petite set C, functions $f : \mathbb{E} \to [1, \infty)$, $g : \mathbb{E} \to \{1, 2, 3, \ldots\}$, $h : \mathbb{E} \to (0, \infty)$, and constants $A > 0$ and $\kappa > 1$ that satisfy the following:*

$$\begin{aligned} E_x[\kappa^{h(x)} f(X_{g(x)})] &\leq f(x), \qquad x \notin C \\ \sup_{x \in C} E_x[\kappa^{h(x)} f(X_{g(x)})] &< \infty \\ \sup_{x \in C} g(x) &< \infty \\ g(x) &\leq Ah(x), \qquad x \notin C \end{aligned}$$

Proof The proof of this result is very similar to the proofs of Theorems 5 and 6. ∎

Remark 1 Theorems 3, 5, and 6 should allow us in principle to use tools from optimal stopping theory (see Shiryayev [30]) to determine whether or not various Markov chains are stable. In [30], the following examples were studied in detail:

$$\begin{aligned} s(x) &= \inf_{\tau} E_x[f(X_\tau)], \\ s(x) &= \inf_{\tau} E_x[f(X_\tau) + \tau]. \end{aligned}$$

If we could guarantee the existence of an optimal stopping time τ^*, then we would only have to show that $\{x : P_x(\tau^* = 0) > 0\}$ is petite. To do this, we would have to

find a nontrivial upper bound h of the payoff function s and show that $h(x) < f(x)$ for all x outside of a petite set.

Remark 2 It is well known that the converse to Theorem 3.2 is also true (let $\tau = 1$, and consult chapter 11 of [21]). Analogously, here is a converse to Theorem 3.

Proposition 8 *If the Markov chain X is Harris recurrent, then there exist an $f : \mathbb{E} \to R_+$ unbounded off petite sets, a finite stopping time $\tau \geq 1$, and a petite set C such that for $x \notin C$,*

$$E_x[f(X_\tau)] \leq f(x).$$

Proof By ψ-irreducibility, we know from Theorem 5.5.5 of [21] that there exists an increasing sequence of sets $\{C_n\}$ where C_n is petite for each n, and $\cup_{n=1}^{\infty} C_n = \mathbb{E}$. Let r be a large enough integer such that $\psi(C_r) > 0$. Then $\tau_{C_r} < \infty$ almost surely, for any initial starting point $x \in \mathbb{E}$.

Now consider the real-valued function $f(x) := \min\{n \geq 1 : x \in C_n\}, \qquad x \in \mathbb{E}$. Clearly f is unbounded off petite sets. Moreover, $E_x(f(X_{\tau_{C_r}})) \leq r < f(x), \qquad x \notin C_r$.

∎

CHAPTER III

USING PALM MEASURES TO ANALYZE TRANSIENT PROPERTIES OF QUEUEING SYSTEMS

3.1 Introduction

It is often of interest to analyze queues from the perspective of an arriving or departing customer. An early example of this comes from the analysis of the $M/G/1$ queue: one can use elementary Markov chain theory to study the embedded queue-length process at the departure times. The stationary distribution of this embedded process gives us the long-run fraction of departing customers that observe the system being in a particular state.

However, it is also of interest to determine the likelihood that other, more complicated events occur, while still observing things from the perspective of the departing customer in steady-state. To do this, one uses a Palm measure. This measure is induced by a stationary version of the point process on the doubly infinite time axis, and under the Palm measure, a point is placed at time 0 with probability one. Since our process is stationary, the customer at time 0 can be considered to be in steady-state.

Many classical results from queueing theory can be stated in terms of Palm measures induced by stationary systems. For instance, one can use these measures to

state and prove *ASTA* (arrivals see time averages) results (see [5]; for a survey of *ASTA*, see [20]). A well-known special case is the *PASTA* theorem, which says that in steady-state, the number of customers observed by an arrival is equal in distribution to the number of customers in the system at an arbitrary point of time. This particular result was first proved rigorously by Wolff [36], but it was "known" long before this. For instance, Kleinrock [17] provides an intuitive proof of *PASTA* by first arguing intuitively that it holds in a transient sense.

Palm measures can also be used to prove expectation versions of Little's law, which says that in steady state, the expected number of customers (with respect to the underlying measure) in the system is equal to the arrival rate of customers, times the expected amount of time (with respect to the Palm measure) a customer that arrives at time 0 spends in the system.

Our focus is on viewing queues from the perspective of a customer that arrives at time t, without assuming that steady-state has been reached. To do this, one needs to consider an entire family of Palm measures, indexed by the real numbers. A discussion of such measures can be found in such references as [11, 16]. We first show that, even in the non-stationary setting, there is a connection between the stochastic intensity of a point process and the family of Palm measures induced by the same point process. Papangelou first proved this result for stationary point processes, and Bremaud later generalized the result for use in his study of stationary queueing systems. We then, as a corollary to our result, provide necessary and sufficient conditions that the family of Palm measures must satisfy for a point process to be Poisson (with respect to a filtration). This is a non-stationary generalization of what Bremaud refers to as

Mecke's characterization of a Poisson process.

The motivation behind Bremaud's generalization of Papangelou's result was to generate necessary and sufficient conditions for the $ASTA$ property to hold. Similarly, we show that essentially the same conditions are required for an $ASTA$-like property to hold at a time t (so in our case, we are not looking at a time-average). This shows that $PASTA$ and Conditional $PASTA$ holds in a transient sense, for almost every t.

Our next collection of results include limit lemmas associated with Palm probabilities. The first lemma gives some insight on how probabilities of the form $P_t(X(t) \in A)$ behave for large t. The second lemma shows how to approximate probabilities of the form $P_t(X(t) \in A)$. One interesting consequence of our limit theorems is that they provide the necessary justification needed to make Kleinrock's $PASTA$ proof rigorous.

In the next section, we show that Palm probabilities can be simplified when our processes exhibit a semi-regenerative structure. This is used to derive, for example, the Laplace transform of the Laplace transform of the waiting time experienced by a customer at time t (this double transform is similar to the double transform of the queue length) of an $M/M/1$ $FIFO$ queue. We then focus on Palm probabilities induced by point processes that consist of points that are a subset of the transition times of a continuous-time Markov chain. These probabilities can be considerably simplified in terms of our underlying measure, due to both the existence of a stochastic intensity and the existence of a semi-regenerative phenomenon.

We conclude by showing how different forms of Little's law carry over to the non-stationary setting. Our results are very similar to the results of [4], but our approach is much simpler, in that we don't need to assume that various limits exist. We also

show that the use of Palm probabilities allows one to derive relationships between any moment of the queue length at time t, to the waiting times of customers that arrive in the interval $(0, t]$.

3.2 Palm Measures

We will use the following notation throughout the paper. Let $\{N(B) : B \in \mathcal{B}\}$ denote a point process on $\Re$ defined on a probability space $(\Omega, \mathcal{F}, P)$, where $N(B)$ is the number of points in B and $\mathcal{B}$ is the family of Borel sets in $\Re$. That is,

$$N(B) = \sum_n \mathbf{1}(T_n \in B), \qquad B \in \mathcal{B},$$

where $\ldots \leq T_{-2} \leq T_{-1} \leq T_0 \leq 0 < T_1 \leq T_2 \leq \ldots$ are the point locations. Assume the mean measure $\mu(B) = E[N(B)]$ is σ-finite.

The main focus of our study are time-dependent Palm probabilities of N. Such measures require some structure on the probability space. Accordingly, we assume throughout that Ω is a complete, separable metric space, and $\mathcal{F}$ is its associated Borel sets. Under the preceding assumptions, there exists a (μ-a.e. unique) probability kernel $P_t(A)$ such that

$$E[N(B)\mathbf{1}(\omega \in A)] = \int_B P_t(A)\mu(dt), \qquad A \in \mathcal{F}, B \in \mathcal{B}. \tag{13}$$

This is proved in [11, 16]. Here $\mathbf{1}(\textit{statement})$ is the indicator function that is 1 or 0 according as the "statement" is true or false.

Definition 9 The collection P_t for μ-a.e. t defined by (13) is the family of time-dependent *Palm probabilities* induced by the point process N. We interpret $P_t(A)$

as the probability of A, given that N has a point at time t, and let E_t denote the expectation under P_t.

A major tool for the analysis below involving Palm probabilities is the following Campbell-Mecke formula , which follows from (13).

Theorem 10 *For any measurable* $f : \Re \times \Omega \to \Re_+$,

$$\int_\Omega \int_\Re f(t,\omega) N(dt) P(d\omega) = \int_\Re \int_\Omega f(t,\omega) P_t(d\omega) \mu(dt). \tag{14}$$

Associated with the point process N, we will consider a stochastic process $\{X(t) : t \in \Re\}$ defined on $(\Omega, \mathcal{F}, P)$ that takes values in a complete, separable metric space $\mathbb{E}$. We assume $X(t)$ is a measurable process and its paths are right-continuous with left-hand limits. The dependency between N and X will be specified later in the various settings. Note that for real-valued $X(t)$, the Campbell-Mecke formula is

$$E[\int_\Re X(t) N(dt)] = \int_\Re E_t[X(t)] \mu(dt), \tag{15}$$

provided the expectations exist. Then by the definition of Radon-Nikodym derivatives,

$$E_t[X(t)] = \frac{E[X(t)N(dt)]}{\mu(dt)}. \tag{16}$$

The theory of Palm probabilities for stationary point processes is well understood; see for instance [3]. However, there are only a few studies dealing with time-dependent Palm probabilities for non-stationary processes [16], and even fewer studies are related to queueing [26]. We will now show a relation between time-dependent Palm probabilities and a Palm probability for a stationary system.

Suppose for now that N is stationary (i.e., its increments are stationary). That is, $S^t N \stackrel{d}{=} N$, $t \in \Re$, where S^t is the time-shift operator defined by $S^t N = \{N(B+t) : B \in \mathcal{B}\}$. With no loss in generality, assume

$$S^t N(\omega, B) = N(\theta_t \omega, B), \qquad \omega \in \Omega,\ t \in \Re,\ B \in \mathcal{B},$$

where θ_t is a stationary process (called a flow) on Ω that satisfies

$$\theta_s(\theta_t(\omega)) = \theta_{s+t}(\omega), \qquad \theta_0(\omega) = \omega.$$

Also, assume N is simple (its point locations are distinct a.s.) and $\lambda = E[N(0,1]]$ is finite. Then $\mu(B) = E[N(B)] = \lambda \int_B dt$, and λ is the (constant) *intensity* of N.

Under these assumptions, there is a "single" Palm probability P^0 for N conditioned that it has a point at 0, which is defined by

$$P^0(A) = \lambda^{-1} E[\int_0^1 \mathbf{1}(\theta_t \in A) dt]. \tag{17}$$

It is well-known that $P^0\{N(\{0\}) = 1\} = 1$ (i.e., N has a point at 0 a.s. under P^0), and the Campbell-Mecke formula is

$$E\Big[\int_{\Re} f(t, \theta_t) N(dt)\Big] = \lambda \int_{\Re} E^0[f(t,\omega)] dt. \tag{18}$$

The next result shows that time-dependent Palm probabilities P_t are invariant in time in the sense that each $P_t \circ \theta_t^{-1}$ is equal to P^0. Here $\mathcal{L}$ denotes Lebesgue measure.

Proposition 11 *For the simple stationary point process N described above,*

$$P_t \circ \theta_t^{-1} = P^0, \quad \textit{for } \mathcal{L}\textit{-a.e. } t.$$

Proof For any $A \in \mathcal{F}$ and $B \in \mathcal{B}$, expressions (15) and (18) yield the respective equations

$$\begin{aligned} E\Big[\int_B \mathbf{1}(\theta_t \in A) N(dt)\Big] &= \lambda \int_B P_t(\theta_t \in A) dt, \\ E\Big[\int_B \mathbf{1}(\theta_t \in A) N(dt)\Big] &= \lambda \int_B P^0(A) dt. \end{aligned}$$

Thus, it follows that $P_t(\theta_t \in A) = P^0(A)$ for a.e. t, which proves the assertion. ∎

The preceding terminology and results also apply to processes defined on the nonnegative time axis $\Re_+$. In this setting we let $N(t)$ and $N(a, b]$ denote the number of points in the respective intervals $(0, t]$ and $(a, b]$.

3.3 Palm Probabilities and Stochastic Intensities

In this section, we present a relationship between the stochastic intensity of a point process and the Palm measures induced by the point process. This is a non-stationary version of Brémaud's result (who refers to his theorem as Papangelou's theorem). We then use this to provide a non-stationary analogue of Mecke's characterization of a Poisson process.

Suppose that N and X are processes defined on the nonnegative time axis as above, but with the additional property that they are adapted to a filtration $\mathcal{F}_t$, $t \geq 0$. Assume, for each rational number s, that $\mathcal{F}_s$ is countably generated by a π-system $\mathcal{C}_s$ (a collection of subsets that is closed under finite intersections — if two finite measures agree on a π-system, they agree on the σ-field generated by that π-system [15]). This assumption is automatically satisfied by the filtration generated

by (X, N).

We will use the following terminology for functions. Suppose $f : \Re_+ \times \Omega \to \Re$ is a measurable function. The f is $\mathcal{F}_t$-*adapted* if $f(t, \cdot)$ is $\mathcal{F}_t$-measurable, for each t. The f is $\mathcal{F}_t$-*progressive* if, for any fixed $t \in \Re_+$, the set $\{(s, \omega) \in [0, t] \times \Omega : f(s, w) \in A\} \in \mathcal{B}[0, t] \bigotimes \mathcal{F}_t$. Finally, f is $\mathcal{F}_t$-*predictable* if $f(0, \cdot)$ is $\mathcal{F}_0$-measurable, and

$$\{(t, \omega) \in (0, \infty) \times \Omega : f(t, \omega) \in A\} \in \mathcal{P}(\mathcal{F}_t),$$

where $\mathcal{P}(\mathcal{F}_t)$ is the σ-field generated by the rectangles $(s, t] \times B$, where $0 \leq s \leq t$ and $B \in \mathcal{F}_s$.

We will consider a time-dependent randomized intensity for N based on the information contained in the $\mathcal{F}_t$.

Definition 12 A $\mathcal{F}_t$-progressive function $\lambda : \Re_+ \times \Omega \to \Re_+$ is a $\mathcal{F}_t$-*intensity of* N (under P) if

$$E[N(a, b]|\mathcal{F}_a] = E\Big[\int_{(a,b]} \lambda(t)dt\Big|\mathcal{F}_a\Big], \qquad (a, b] \in \mathcal{B}.$$

Here is a key property of an intensity [6]; it is analogous to the Campbell-Mecke formula.

Theorem 13 *If λ is an $\mathcal{F}_t$- intensity of N and $Y(t)$ is a nonnegative predictable process, then*

$$E\Big[\int_{\Re_+} Y(t)N(dt)\Big] = \int_{\Re_+} E[Y(t)\lambda(t)]dt. \tag{19}$$

We will now present necessary and sufficient conditions for $P_t << P$ on $\mathcal{F}_{t-}$, for μ-a.e. t. This is a non-stationary generalization of Papangelou's result, which states that for a stationary point process N on the real line, $P^0 << P$ on $\mathcal{F}_{0-}$ if and only if

the point process has an $\mathcal{F}_t$-intensity. Brémaud later extended this result to histories that aren't necessarily the history induced by the point process. His proof includes a lemma that all predictable processes on the line have a nice form, and he uses this form along with the θ_t-invariance of P to prove the result.

Our approach is different. We show that when enough σ-fields are countably generated, a well-known martingale approximation of Radon-Nikodym derivatives yields an intensity that's $\mathcal{F}_t$-predictable.

Theorem 14 *For the point process N (recall μ is its mean measure), the following statements are equivalent:*

(a) *N has an $\mathcal{F}_t$-intensity.*

(b) *$P_t << P$ on $\mathcal{F}_{t-}$, μ-a.e. t and $\mu << \mathcal{L}$.*

The following proof shows how to construct a $\mathcal{F}_t$-intensity λ that is $\mathcal{F}_t$-predictable, when $P_t << P$ on $\mathcal{F}_{t-}$. Furthermore, the Palm probabilities are related to the intensity by

$$\frac{dP_t}{dP} = \frac{\lambda(t)}{E[\lambda(t)]}.$$

Consequently, for the process $X(t)$ described above, and bounded continuous $f : \mathbb{E} \to \Re$,

$$E_t[f(X(t-))] = \frac{E[f(X(t-))\lambda(t)]}{E[\lambda(t)]}. \tag{20}$$

This formula is the basis of "arrivals see time averages" in the next section.

Proof (a) $\Rightarrow$ (b). If N has a $\mathcal{F}_t$-intensity λ, then clearly $\mu(B) = \int_B E[\lambda(t)]dt$, and so

$\mu << \mathcal{L}$. Next, fix $A \in \mathcal{F}$ and define $t_A = \inf\{t : A \in \mathcal{F}_t\}$ and

$$Y^A(t,\omega) = \mathbf{1}(\omega \in A)\mathbf{1}(t > t_A).$$

It is clear that Y^A is predictable. Therefore, by the Campbell-Mecke formula and (19),

$$\begin{aligned} \int_B E_t[Y^A(t)]\mu(dt) &= E[\int_B Y^A(t)N(dt)] \\ &= \int_B E[Y^A(t)\lambda(t)]dt, \qquad B \in \mathcal{B}. \end{aligned}$$

From this and $\mu(dt) = E[\lambda(t)]dt$, it follows that, for any set A in our countable separating class,

$$E_t[Y^A(t)] = \frac{E[Y^A(t)\lambda(t)]}{E[\lambda(t)]}, \qquad \mu\text{-a.e. t.}$$

Fix a t in the set of points that satisfy this equality for all sets contained in our separating class. This class forms a π-system, so we see that, for any rational $s < t$,

$$P_t(A) = \int_A \frac{\lambda(t)}{E[\lambda(t)]}dP, \qquad A \in \mathcal{F}_s.$$

By using the same type of monotone-class argument, we see that this expression is true for all $A \in \mathcal{F}_{t-}$. This proves $P_t << P$ on $\mathcal{F}_{t-}$, μ-a.e. t, which finishes the proof that (a) $\Rightarrow$ (b).

(b) $\Rightarrow$ (a) Assume (b) is true, and let h denote the density of μ. Then by the Campbell-Mecke formula,

$$\int_A N(a,b]dP = E\big[N(a,b]\mathbf{1}(\omega \in A)\big] = \int_{(a,b]} P_t(A)h(t)dt, \qquad A \in \mathcal{F}_a.$$

Suppose for now that there is a $\mathcal{F}_t$-progressive nonnegative function $f(t,w)$ such that

$$P_t(A) = \int_A f(t,\omega)P(d\omega). \tag{21}$$

Then from above,

$$\begin{aligned}\int_A N(a,b]dP &= \int_A \int_{(a,b]} f(t,\omega)h(t)dtP(d\omega) \\ &= \int_A E[\int_{(a,b]} f(t,\omega)h(t)dt|\mathcal{F}_a]P(d\omega).\end{aligned}$$

Hence, N has a stochastic intensity.

It remains to define a $\mathcal{F}_t$-progressive nonnegative function $f(t,w)$ that satisfies (21). Consider the function

$$\begin{aligned}f(t,\omega) = \limsup_{n\to\infty} \sum_{m=1}^{r_{0,n}} \frac{P_t(B^n_{0,m})}{P(B^n_{0,m})}\mathbf{1}(\omega \in B^n_{0,m})\mathbf{1}(t \in [0, \frac{1}{2^n}]) \\ + \sum_{k=1}^{2^{2n}} \sum_{m=1}^{r_{k,n}} \frac{P_t(B^n_{k,m})}{P(B^n_{k,m})}\mathbf{1}(\omega \in B^n_{k,m})\mathbf{1}(t \in (\frac{k}{2^n}, \frac{k+1}{2^n}]),\end{aligned}$$

where $\{B^n_{k,m}\}_m$ is the $(k,n)^{th}$ finite partition of Ω that consists of $\mathcal{F}_{\frac{k}{2^n}}$-measurable sets, and $r_{k,n}$ represents the number of sets in the $(k,n)^{th}$ partition that have positive P-measure. We assume that for any fixed dyadic rational $\frac{k_0}{2^{n_0}}$, the sequence of partitions $\{B^{n_0+p}_{k_0 2^p,m}\}_m$ becomes finer and finer as p increases, and it also generates $\mathcal{F}_{\frac{k_0}{2^{n_0}}}$. We further refine our partitions so that $\{B^n_{k,m}\}_m$ becomes finer and finer as k increases (for fixed n). Notice that f is $\mathcal{F}_t$-predictable, which implies that it is also $\mathcal{F}_t$-progressive. Finally, we can conclude from a martingale approximation of Radon-Nikodym derivatives (see Application (VIII) of Section 9.5 of [8]) that for each t, $f(t)$ is the Radon-Nikodym derivative of P_t with respect to P on $\mathcal{F}_{t-}$, since for any fixed t, the resulting sequence of partitions must generate $\mathcal{F}_{t-}$. To see this, notice that for any fixed t, we see that the corresponding sequence of partitions generate $\mathcal{F}_s$, for every dyadic rational $s < t$, which implies that the sequence also generates $\mathcal{F}_{t-}$. Thus, f satisfies (21). ∎

Suppose for the moment that N is a stationary point process on $\Re$. Often, it may be of interest to know whether or not N is a $\mathcal{F}_t$-Poisson process.

Definition 15 A point process N is a *$\mathcal{F}_t$-Poisson process* with deterministic rate function $\lambda(t)$ if, for any $a \leq b$, and $k \geq 0$,

$$P(N(a,b] = k|\mathcal{F}_a) = e^{-\mu(a,b]}(\mu(a,b])^k/k!,$$

where $\mu(a,b] = \int_a^b \lambda(t)dt$.

Mecke showed that a stationary point process N is $\mathcal{F}_t$-Poisson if and only if $P^0 = P$ on $\mathcal{F}_{0-}$. The proof of the result is a simple consequence of Brémaud's version of Papangelou's theorem.

Our Papangelou-type theorem suggests that a similar criterion should exist for non-stationary point processes. This is due to the fact that the structural relationship between P_t and P on $\mathcal{F}_{t-}$ is essentially the same as that between P^0 and P on $\mathcal{F}_{0-}$ in the stationary context.

Theorem 16 *The point process N is a $\mathcal{F}_t$-Poisson process with deterministic rate function λ if and only if $P_t = P$ on $\mathcal{F}_{t-}$, μ-a.e. t and $\mu << \mathcal{L}$.*

Proof Suppose N is $\mathcal{F}_t$-Poisson. Then its $\mathcal{F}_t$-intensity is just its rate function λ, and from the proof above, for $A \in \mathcal{F}_{t-}$,

$$P_t(A) = \int_A \frac{\lambda(t)}{E[\lambda(t)]} dP = P(A).$$

Now assume the converse is true. For $A \in \mathcal{F}_a$,

$$\begin{aligned}\int_A E[N(a,b]|\mathcal{F}_a]dP &= \int_{(a,b]} P_t(A)\mu(dt) \\ &= P(A)\mu(a,b] = \int_A \int_{(a,b]} \lambda(t)dtdP.\end{aligned}$$

This proves that $N(t) - \int_{(0,t]} \lambda(s)ds$ is a $\mathcal{F}_t$-martingale, and hence N is a $\mathcal{F}_t$-Poisson process with rate function λ by Watanabe's theorem below. ∎

Theorem 17 (Watanabe) *If λ is a locally integrable nonnegative measurable function such that $N(t) - \int_{(0,t]} \lambda(s)ds$ is a $\mathcal{F}_t$-martingale, then N is a $\mathcal{F}_t$-Poisson process with rate function λ.*

We will see in the next section that our Palm characterization of $\mathcal{F}_t$-Poisson processes leads to a quick proof of transient *PASTA* phenomena.

3.4 Arrivals See Time Averages

Consider the processes N and X as in the preceding section, where $\mathcal{F}_t$, $t \in \Re_+$ is the natural filtration generated by (N, X). Suppose N has a $\mathcal{F}_t$-intensity λ.

In the Palm context, the *ASTA* problem is typically the problem of determining whether or not $P(X(0-) \in A) = P^0(X(0-) \in A)$, where P^0 is the Palm probability induced by N and (N, X) are jointly stationary. Melamed and Whitt [19] focus on the slightly different issue of finding conditions under which the quantities

$$\overline{V}(t) = \frac{E\left[\int_0^t f(X(s-))r(s)ds\right]}{\mu(t)},$$

$$\overline{W}(t) = \frac{E\left[\int_0^t f(X(s-))N(ds)\right]}{E[N(t)]}$$

converge to the same limit. Here $r(t) = E[\lambda(t)]$, and $\mu(t) = \int_0^t r(s)ds$, where $\lambda(t)$ is the stochastic intensity of N. Theorem 1 in [19] says that if f is bounded and continuous,

$$\overline{W}(t) = \frac{\int_0^t E\left[f(X(s-))\lambda(s)\right] ds}{\mu(t)} = \overline{V}(t) + \frac{\int_0^t Cov(f(X(s)), \lambda(s))ds}{\mu(t)}.$$

Now when the limits $\overline{W}(\infty) = \lim_{t\to\infty} \overline{W}(t)$, $\overline{V}(t) = \lim_{t\to\infty} \overline{V}(t)$, and $r(\infty) = \lim_{t\to\infty} t^{-1}\mu(t)$ exist, then

$$r(\infty)(\overline{W}(\infty) - \overline{V}(\infty)) = \lim_{t\to\infty} t^{-1} \int_0^t Cov(f(X(s)), \lambda(s))ds.$$

Therefore, we have *ASTA* if the covariance term on the right approaches zero in the limit. In the next section, we will show that $\overline{W}(\infty) = \overline{V}(\infty)$ implies that *ASTA* holds, if we assume that a jointly stationary version of (N, X) can be constructed. For a good overview of the *ASTA* literature, see [20].

Furthermore, it was also established in [5] (and also in [19] by different methods) that if (N, X) is jointly stationary, then

$$E^0[f(X(0-))] = \frac{E[f(X(0-))\lambda(0)]}{E[\lambda(0)]}.$$

This implies that we have *ASTA* if and only if $E[\lambda(0)|X(0-)] = E[\lambda(0)]$, i.e. if the stationary process has a Lack of Bias at time 0.

With these results in mind, our Papangelou-type theorem suggests that the *ASTA* phenomenon exists at a fixed time t, if and only if we have a Lack of Bias (23) at t.

Theorem 18 (*ASTA*) *Suppose t is such that $P_t << P$ on $\mathcal{F}_{t-}$. Under the preceding assumptions,*

$$E_t[f(X(t-))] = E[f(X(t-))], \tag{22}$$

for each bounded continuous function $f : \mathbb{E} \to \Re_+$*, if and only if*

$$E[\lambda(t)|X(t-)] = E[\lambda(t)]. \quad \text{(Lack of Bias)} \tag{23}$$

Proof This proof parallels one in Melamed and Whitt [19]. Suppose t is such that $P_t << P$ on $\mathcal{F}_{t-}$. Then from (20),

$$E_t[f(X(t-))] = \frac{E[f(X(t-))\lambda(t)]}{E[\lambda(t)]}.$$

Now, if $E_t[f(X(t-))] = E[f(X(t-))]$, then

$$E[f(X(t-))]E[\lambda(t)] = E[f(X(t-))\lambda(t)].$$

This holds for all bounded continuous f, so it follows that this equality also holds for all bounded measurable f. Thus (23) is true.

Conversely, if (23) is true, then (22) follows since

$$E_t[f(X(t-))] = \frac{E[f(X(t-))\lambda(t)]}{E[\lambda(t)]} = E[f(X(t-))].$$

∎

Now we'll suppose that N is a $\mathcal{F}_t$-Poisson process with deterministic rate function $\lambda(t)$. What's very interesting about this case is that N being $\mathcal{F}_t$-Poisson implies that a *PASTA*-like phenomenon exists within (N, X), regardless of the limiting behavior of X.

Theorem 19 $(PASTA)$ *If* N *is a* $\mathcal{F}_t$*-Poisson process, then*

$$E_t[f(X(t-))] = E[f(X(t-))], \qquad \mu - a.e.\ t.$$

Proof This follows either by Corollary 16 or Theorem 18. ∎

This property intuitively means that the distribution of $X(t-)$ given that N has a point at time t is the same as the distribution of $X(t-)$ without knowing whether or not N has a point at t. In other words, knowing that N has a point at time t does not bias the distribution of $X(t-)$. Of course, when $X(t)$ is continuous in distribution ($X(t-) \stackrel{d}{=} X(t)$), then PASTA says $P_t(X(t-) \in \cdot) = P(X(t) \in \cdot)$.

It should be noted that this property has been conjectured before by Kleinrock [17], who provides an intuitive argument that implicitly assumes (he makes no mention of Palm probabilities) that for any queueing system with Poisson arrivals, we know that

$$P_t(X(t-) = k) \stackrel{?}{=} \lim_{h \downarrow 0} P(X(t-) = k | N[t, t+h) \geq 1), \tag{24}$$

where X is a nonnegative integer-valued queueing process, N is the arrival process of X, and P_t is the Palm probability induced by N. By assuming that (N, X) satisfies a *Lack of Anticipation* assumption (see [36]), we see that

$$\lim_{h \downarrow 0} P(X(t-) = k | N[t, t+h) \geq 1) = P(X(t-) = k). \tag{25}$$

As $t \to \infty$, we expect the left-hand-side of (25) to approach the long-run fraction of time an arrival observes k customers in the system, and we know that the right-hand-side of (25) approaches the long-run fraction of time the system is in state k.

It is not mathematically clear at this point, however, that (24) holds for μ-a.e.t. Moreover, it is also unclear that the right-hand-side of (24) actually converges. In the next section, we will establish a convergence result for probabilities of the form

$P_t(X(t-) = k)$, and we will show when probabilities of this form can be represented as a limit that's similar to the right-hand-side of (24).

Next, we'll show that there is also a transient *PASTA* phenomenon under a certain type of conditioning (the corresponding limit theorem was first noticed in [32]). Suppose N is a point process with a $\mathcal{F}_t$-intensity $g(Y(t))$, where $Y(t)$ is integer-valued and $\mathcal{F}_t$-predictable. Consider the point process N_i on $\Re_+$ defined by

$$N_i(B) = \int_B \mathbf{1}(Y(t) = i)N(dt),$$

where i is such that $P(Y(t) = i) > 0$, $t > 0$. Let P_t^i denote the Palm measures induced by N_i and let E_t^i denote the associated expectation.

Theorem 20 *Under the preceding conditions,*

$$E_t^i[f(X(t-))] = E[f(X(t-))|Y(t) = i], \qquad \mu - a.e.\ t.$$

Proof The function $\lambda : \Re_+ \times \Omega \to \Re_+$ defined by $\lambda(t, \omega) = g(i)\mathbf{1}(Y(t) = i)$ is a stochastic intensity of N_i. It is also predictable, and so by (20), for any bounded continuous f,

$$E_t^i[f(X(t-))] = \frac{E[f(X(t-))g(i)\mathbf{1}(Y(t) = i)]}{g(i)P(Y(t) = i)} = E[f(X(t-))|Y(t) = i].$$

∎

Example 21 ($M/G/c$ Queue) Consider a single-server queueing system with Poisson arrivals of rate λ, and a capacity of size c. If N represents the arrival process

of customers to the system (including those that do not join the system), then the number of customers that join the system in $(0, t]$ is just

$$N^*(t) = \int_0^t \mathbf{1}(X(s-) \leq c-1) N(ds).$$

Furthermore, if X is left-continuous and adapted to a filtration $\mathcal{F}_t$, then it's $\mathcal{F}_t$-predictable, and so for any $\mathcal{F}_t$-predictable process Z

$$E\Big[\int_0^t Z(s)\mathbf{1}(X(s-) \leq c-1) N(ds)\Big] = E\Big[\int_0^t Z(s)\mathbf{1}(X(s-) \leq c-1)\lambda ds\Big]$$

This implies that an $\mathcal{F}_t$-intensity of N^* is $\lambda \mathbf{1}(X(s-) \leq c-1)$. Using Theorem (20), we find that, for a customer that joins the system at a time t, the probability that he/she observes $k \leq c-1$ customers in the system is equal to $P(X(t-) = k | X(t-) \leq c-1)$.

Example 22 (Queues with Markov Modulated Arrivals) Consider a generalization of the queue in the previous example to one that consists of an arrival process of potential customers that is a Markov modulated Poisson process. In this case, we know that the $\mathcal{F}_t$-intensity of such a process is of the form $g(Y(t-))$, where Y is a Markov process with a finite, integer-valued state space. In this case, if we denote N_i^* as the point process of customers that join the system while observing that the process Y is in state i, then

$$\begin{aligned} E\int_0^t Z(s) N_i^*(ds) &= E\int_0^t Z(s)\mathbf{1}(X(s-) \leq c-1)\mathbf{1}(Y(s-) = i) N(ds) \\ &= E\int_0^t Z(s)\mathbf{1}(X(s-) \leq c-1)\mathbf{1}(Y(s-) = i) g(Y(s-)) ds \\ &= E\int_0^t Z(s)\mathbf{1}(X(s-) \leq c-1)\mathbf{1}(Y(s-) = i) g(i) ds. \end{aligned}$$

Therefore, we see that if P_t^i is the arrival process induced by N_i^*, then for almost all

t,

$$P_t^i(X(t-) = k) = P(X(t-) = k | X(t-) \leq c-1, Y(t-) = i).$$

3.5 Limiting Behavior of Palm Probabilities

The first rigorous proof of *PASTA* was given by Wolff [36], who works under the assumption that (N, X) satisfies a Lack of Anticipation property. What's interesting is that Kleinrock also uses this property in his intuitive proof of *PASTA* [17]. He was interested in showing that the limiting probability $p(k) = \lim_{t\to\infty} P(X(t-) = k)$ is also equal to the long-run fraction of time that an arrival to the system observes k customers in the system. To do this, he first shows that for each $t > 0$,

$$P(X(t-) = k) = \lim_{h \downarrow 0} P(X(t-) = k | N[t, t+h) \geq 1) \tag{26}$$

If we let $t \to \infty$ in (26), and assume that the limits exist, we see that

$$p(k) = \lim_{t\to\infty} \lim_{h\downarrow 0} P(X(t-) = k | N[t, t+h) \geq 1).$$

At this point, he claims (without proof) that the right-hand-side of this equation represents the long-run fraction of time that an arrival to the system observes k customers in the system. This should follow from the fact that the right-hand-side of (26) could possibly be interpreted as the probability that $X(t-) = k$, given that an arrival occurred at time t. This, however, needs to be justified.

From what we know about Palm probabilities, we expect that

$$P_t(X(t-) \in A) = \lim_{h\downarrow 0} P(X(t-) \in A | N[t, t+h) \geq 1),$$

In Theorem (24) below, we will establish a similar limit representation for $P_t(X(t-) \in$

A), and this will be a first step towards showing that Kleinrock's intuitive argument is indeed rigorous.

Before this is done, we will first need the following lemma based on the representation (16) of Palm probabilities as Radon-Nikodym derivatives.

Lemma 23 *For μ-a.e.t,*

$$P_t(S^tX \in C) = \lim_{n\to\infty} \frac{E\left[\int_{B_n(t)} \mathbf{1}(S^uX \in C)N(du)\right]}{\mu(B_n(t))]}, \tag{27}$$

where $B_n(t) = (\frac{k(t)}{2^n}, \frac{k(t)+1}{2^n}]$ is the unique interval that contains t.

Proof The idea behind the proof involves approximating a Radon-Nikodym derivative with a martingale; see Application (VIII) of Section 9.5 of [8]. The only difference here is that our measure μ is not a probability measure, but a similar argument goes through in this case because μ is σ-finite. ∎

We will now present a refinement of the limiting formula in Lemma 23. Assume the process X has piecewise-constant sample paths that have at most a finite number of jumps in finite time intervals. Let N denote a point process on $\Re$ of a certain subset of jump times and let M denote the point process of the other jump times. Assume N is simple and that, for each t,

$$\lim_{\epsilon\downarrow 0} E[N(t-\epsilon, t+\epsilon]\mathbf{1}(N(t-\epsilon, t+\epsilon] \geq 2]/P(N(t-\epsilon, t+\epsilon] \geq 1) \to 0. \tag{28}$$

This condition is satisfied by a Poisson process.

As in Lemma 23, let $B_n(t) = (\frac{k(t)}{2^n}, \frac{k(t)+1}{2^n}]$ be the unique interval that contains t, and let $\nu_n(t) = N(B_n(t))$ and $u_n(t) = k(t)/2^n$. As usual P_t refers to the Palm probabilities for N.

Theorem 24 *Suppose the process X described above is such that, for each $t \geq 0$*

$$\lim_{n\to\infty} P(M(B_n(t)) \geq 1 | N(B_n(t)) \geq 1) = 0. \tag{29}$$

Then for μ-a.e.t,

$$P_t(S^t X \in C) = \lim_{n\to\infty} P(S^{u_n(t)} X \in C) | N(B_n(t)) \geq 1). \tag{30}$$

Many queueing processes satisfy the conditions in this theorem. For example, if X is an $M/M/1$ queueing process with arrival rate λ and service rate μ, then

$$\lim_{n\to\infty} P(M(B_n(t)) \geq 1 | N(B_n(t)) \geq 1) \leq \lim_{n\to\infty} 1 - e^{-\frac{\mu}{2^n}} = 0.$$

Proof We will prove (30) by proving the equivalent statement that, for any bounded function $f : D(\Re) \to \Re_+$,

$$E_t[f(S^t X)] = \lim_{n\to\infty} \frac{E\Big[f(S^{u_n(t)} X)\mathbf{1}(\nu_n \geq 1)\Big]}{P(N(B_n(t)) \geq 1)}. \tag{31}$$

For simplicity, let $\nu_n = N(B_n(t))$ and

$$Z_n = \int_{B_n(t)} f(S^u X) N(du).$$

By Lemma 23,

$$E_t[f(S^t X)] = \lim_{n\to\infty} \frac{E[Z_n \mathbf{1}(\nu_n \geq 1)\Big]}{E[\nu_n]}. \tag{32}$$

Since N is a simple Poisson process,

$$P(\nu_n \leq 1) \to 1, \qquad P(\nu_n \geq 2) \to 0.$$

Also, by assumption (28),

$$\frac{E[\nu_n]}{P(\nu \geq 1)} = \frac{P(\nu_n = 1)}{P(\nu_n = 1) + P(\nu_n \geq 2)} + \frac{E[\nu_n \mathbf{1}(\nu_n \geq 2)]}{P(\nu_n \geq 1)} \to 1.$$

Using these observations in (32), we have

$$E_t[f(S^tX)] = \lim_{n\to\infty} \frac{E\Big[f(S^{u_n(t)}X)\mathbf{1}(\nu_n \geq 1)\Big]}{P(\nu_n \geq 1)} + d_n + d'_n, \tag{33}$$

where

$$\begin{aligned} d_n &= \frac{E\Big[(Z_n - f(S^{u_n(t)}X))\mathbf{1}(\nu_n = 1)\Big]}{P(\nu_n \geq 1)} \\ d'_n &= \frac{E\Big[(Z_n - f(S^{u_n(t)}X))\mathbf{1}(\nu_n \geq 2)\Big]}{P(\nu_n \geq 1)}. \end{aligned}$$

Clearly $|Z_n - f(S^{(u_n(t)}X)| \leq 2b\nu_n$, where b is an upper bound on f. Then under assumptions (28) and (29),

$$\begin{aligned} |d_n| &\leq \frac{2P(\nu_n = 1, M(B_n(t)) \geq 1)}{P(\nu_n \geq 1)} \to 0, \\ |d'_n| &= \frac{2bE\Big[\nu_n\mathbf{1}(\nu_n \geq 2)\Big]}{P(\nu_n \geq 1)} \to 0. \end{aligned}$$

Applying these limits to (33) proves (31). ∎

It is also of interest to know if one can say something about the behavior of $P_t(X(t-) = k)$ for large t. Our intuition leads us to expect that $P_t(X(t-) = k)$ should get close to $P^0(X(0-) = k)$ as t gets large, if we were to assume that a stationary version of (N, X) can be constructed on the probability space. In general, for large t we expect that $P_t(X(t-) = k)$ should approximately be $\pi(k)$, where (we assume that $\pi(k)$ exists)

$$\pi(k) = \lim_{n\to\infty} \frac{\sum_{k=1}^{n} \mathbf{1}(X(T_k-) = k)}{n}.$$

In the rest of this section, we address the issue of under what conditions $E_t[f(S^tX)]$ converges a.s.-μ as $t \to \infty$, where $f : D(\Re_+) \to \Re$ is a bounded measurable function

with respect to $\tilde{\mathcal{B}}$, which represents the Borel sets in $D(\Re_+)$. Here $h(t) \to c$ as $t \to \infty$ a.s.-μ means a.s. convergence for t in a set whose complement has μ-measure 0.

Proposition 25 *Suppose there is a constant $r > 0$ such that*

$$\lim_{t\to\infty} t^{-1}N(t) = r \ \ a.s. \quad \text{and} \quad \lim_{t\to\infty} t^{-1}E[N(t)] = r.$$

For a bounded $f : D(\Re_+) \to \Re$, if $\alpha = \lim_{n\to\infty} n^{-1}\sum_{k=1}^{n} f(S^{T_k}X)$ exists a.s., then

$$\lim_{t\to\infty} t^{-1}\int_0^t E_u[f(S^uX)]\mu(du) = r\alpha.$$

Proof From the Campbell-Mecke formula and a generalization of the Dominated Convergence theorem, it follows that

$$\begin{aligned}
\lim_{t\to\infty} \frac{1}{t}\int_0^t E_u[f(S^uX)]\mu(du) &= \lim_{t\to\infty} \frac{1}{t}E[\int_0^t f(S^uX)N(du)] \\
&= \lim_{t\to\infty} E[(t^{-1}N(t))N(t)^{-1}\sum_{n=1}^{N(t)} f(S^{T_n}X)] \\
&= r\alpha
\end{aligned}$$

∎

While this lemma doesn't establish whether or not $E_t[f(S^tX)]$ converges μ-a.e. t, it does establish that if it converges, it must converge to α.

Theorem 26 *Let X denote the queue-length process of a $GI/G/1$ FIFO queue, and let N denote its arrival process. Suppose the joint process (X, N) regenerates at the beginning of each busy period, and the inter-arrival time distribution that has a density that's directly Riemann integrable. Then*

$$\lim_{n\to\infty} P_t(X(t-) = k) = P^0(\overline{X}(0-) = k) \quad \mathcal{L}\text{-a.s.} \tag{34}$$

Proof We know from [6] that the $\mathcal{F}_t$-intensity is of the form $h(A(t))$, where h is the hazard function of the interarrival times, and $A(t) = t - T_{N(t)}$ is the age process. Using a renewal argument, it is easy to see that

$$\lim_{t\to\infty} E[h(A(t))] = r.$$

Moreover, we know that for $\mathcal{L}$-a.e. t,

$$P_t(X(t-) = k) = \frac{E[\mathbf{1}(X(t-) = k)h(A(t))]}{E[h(A(t))]}.$$

Since (X, N) regenerates at the beginning of each busy period, it's clear that the term on the right converges. Therefore Proposition 25 yields (34). ∎

Remark 27 Our transient *PASTA* result tells us that $P_t(X(t-) \in A) = P(X(t-) \in A)$ for μ-a.e. t. As t gets large, we see that $P(X(t-) \in A)$ converges to $p(A)$, where p is the limiting distribution of X. Furthermore, if

$$\pi(A) = \frac{\sum_{k=1}^{n} \mathbf{1}(X(T_k-) \in A)}{n},$$

we see from (25) that $P_t(X(t-) \in A)$ converges μ-a.e. t. to $\pi(A)$. Therefore, $p(A) = \pi(A)$, so we have proven a limiting version of *PASTA*, and our argument involves the same type of reasoning that's presented in Kleinrock's argument.

3.6 Palm Prob. for Semi-Regenerative Processes

As above, N and X will denote $\mathcal{F}_t$-adapted processes on the entire time axis $\Re$ and P_t denote the Palm probabilities induced by N. The process X has sample paths in the Skorohod space $D(\Re)$ with Borel σ-field $\tilde{\mathcal{B}}$.

We've seen how probabilities of the form $P_t(S^t X \in C)$ can be expressed in terms of the underlying measure when $\mathbf{1}(S^t X \in C)$ is $\mathcal{F}_t$-predictable. In this section, we will investigate how probabilities of this form can be simplified when X is semi-regenerative with respect to some point process N.

Definition 28 *The process X is* semi-regenerative *with respect to N (whose points T_n are $\mathcal{F}_t$-stopping times) if, for any negative or nonnegative integer n,*

$$P(S^{T_n} X \in C | \mathcal{F}_{T_n}) = p(X(T_n), C), \qquad C \in \tilde{\mathcal{B}}, \tag{35}$$

where $p(x, C)$ is a probability kernel from $\mathbb{E}$ to $D(\Re_+)$.

Note that (35) is essentially a Markov property for $Y_n = S^{T_{n-1}} X$ for $n \geq 1$, but Y_n is not a Markov chain with respect to $\mathcal{F}_{T_n}$ since Y_n need not be $\mathcal{F}_{T_n}$-measurable. There are numerous examples of semi-regenerative processes. For instance, a continuous-time Markov chain is semi-regenerative with respect to its point process of transition times. The queue-length process of a $GI/M/1$ queue is semi-regenerative with respect to its point process of arrivals. This is due to the fact that when a new interarrival time begins, the time until the next potential service completion is also exponential.

Proposition 29 *If X is semi-regenerative with respect to N, then, for μ-a.e. t*

$$P_t(S^t X \in C) = \int_{\mathbb{E}} p(x, C) P_t(X(t) \in dx), \qquad C \in \tilde{\mathcal{B}}.$$

Proof For $B \in \mathcal{B}$,

$$\int_B P_t(S^t X \in C) \mu(dt) \;\; = \;\; E\left[\int_B \mathbf{1}(S^t X \in C) N(dt) \right]$$

$$
\begin{aligned}
&= E\left[\sum_n P(S^{T_n}X \in C, T_n \in B|\mathcal{F}_{T_n})\right] \\
&= E\left[\sum_n p(X(T_n), C)\mathbf{1}(T_n \in B)\right] \\
&= \int_B E_t[p(X(t), C)]\mu(dt) \\
&= \int_B \int_{\mathbb{E}} p(x, C)P_t(X(t) \in dx)\mu(dt).
\end{aligned}
$$

The first equality follows from the Campbell-Mecke formula, and the third follows from the fact that for each n, T_n is $\mathcal{F}_{T_n}$-measurable. ∎

Example 30 Suppose that X is the queue-length process of an $M/M/1$ $FIFO$ queue, and suppose N represents the point process of arrivals. Let $W(t)$ denote the waiting time of the last customer to enter the system at or before time t. Then, for μ-a.e. t,

$$E_t[W(t)] = \frac{E[X(t)] + 1}{\mu}. \tag{36}$$

To see this, note that $W(t) = f(S^t X)$, where $f : D(\Re) \to \Re_+$ is defined by

$$f(x) = \inf\{s > g(x) : \sum_{g(x) \le u \le s} \mathbf{1}(x(u) = x(u-) - 1) \ge x(g(x))\} - g(x)$$

and $g(x) = \sup\{s \le 0 : x(s) = x(s-) + 1\}$. By Proposition 29, for μ-a.e. t,

$$
\begin{aligned}
E_t[W(t)] &= \sum_{n=0}^{\infty} E[W(0)|X(0) = n]P_t(X(t) = n) \\
&= \sum_{n=0}^{\infty} \frac{n}{\mu} P_t(X(t) = n) \\
&= \frac{E_t[X(t)]}{\mu}.
\end{aligned}
$$

Also, $P_t(N(t) = 1) = 1$, so from our transient $PASTA$ result, we see that

$$E_t[X(t)] = E_t[X(t-)] + 1 = E[X(t-)] + 1 = E[X(t)] + 1.$$

Combining the preceding observations proves (36).

We can also say something about the Laplace transform of $W(t)$ under P_t. For $0 < s < 1$,

$$\begin{aligned} E_t[e^{-sW(t)}] &= \sum_{n=1}^{\infty} E_t[e^{-sW(t)}|X(t)=n]P_t(X(t)=n) \\ &= \sum_{n=1}^{\infty} \left(\frac{\mu}{\mu+s}\right)^n P_t(X(t)=n) \\ &= \sum_{n=1}^{\infty} \left(\frac{\mu}{\mu+s}\right)^n P(X(t-)=n-1) \\ &= \frac{\mu}{\mu+s} E\left[\left(\frac{\mu}{\mu+s}\right)^{X(t)}\right]. \end{aligned}$$

This Laplace transform of the transient waiting time of a customer that arrives at time t is as complicated as the generating function of the transient queue length at time t, which has been studied in the literature. However, we can compute the Laplace transform of the generating function of the queue length (see [1]), which means we can also compute the Laplace transform of the waiting time transform. Indeed, assuming $X(0) = i \geq 0$,

$$\begin{aligned} \int_0^{\infty} e^{-\theta t} E_t[e^{-uW(t)}]dt &= \int_0^{\infty} e^{-\theta t} \left(\frac{\mu}{\mu+u}\right) E\left[\left(\frac{\mu}{\mu+u}\right)^{X(t)}\right] dt \\ &= \left(\frac{\mu}{\mu+u}\right) \left[\frac{\left(\frac{\mu}{\mu+u}\right)^{i+1} - \left(\frac{u}{\mu+u}\right)\left(\frac{\xi^{i+1}}{1-\xi}\right)}{\lambda\left(\frac{\mu}{\mu+u} - \xi\right)\left(\eta - \frac{\mu}{\mu+u}\right)}\right]. \end{aligned}$$

where ξ and η are the two roots of the equation

$$\lambda z^2 - (\lambda + \mu + \theta)z + \mu = 0.$$

3.7 Palm Probabilities for Markov Processes

We have seen that the semi-regenerative property allows us to simplify expressions involving Palm probabilities of events that occur in the "future", with respect to the

index of the Palm probability. In this section we present more detailed relationships between P_t and P for the processes X and N as in the preceding section when X is a Markov process. Our Markovian assumption will allow us to derive more detailed relationships by using the stochastic intensity of N to simplify past events, and using the strong Markov property of X to simplify future events.

Throughout this section, we assume X is a Markov jump process on the time axis $\Re$ with transition kernel $q(x, A)$, and M is the point process of its jumps. Let $X_n = X(T_n)$ and $\mathcal{F}_n = \mathcal{F}_{T_n}$. Being a Markov jump process means the sequence $(X_n, T_n - T_{n-1})$ is a $\mathcal{F}_n$-Markov chain with transition probabilities

$$P(X_{n+1} \in A, T_{n+1} - T_n > t | \mathcal{F}_n) = q(X_n, A) e^{-tq(X_n)}, \tag{37}$$

where $q(x) = q(x, \mathbb{E})$. From this it follows that X_n is an $\mathcal{F}_n$-Markov chain with

$$P(X_{n+1} \in A | \mathcal{F}_n) = \frac{q(X_n, A)}{q(X_n)}, \tag{38}$$

and

$$E[T_{n+1} - T_n > t | \mathcal{F}_n] = 1/q(X_n). \tag{39}$$

We will be interested in describing Palm probabilities induced by point processes that consist of points that form a subset of the transition times of the underlying Markov process. Moreover, we will be interested in what we refer to as Palm probabilities induced by *C-events*.

We will use the following terminology from [29].

Definition 31 *A* C-event *of* X *occurs at time* t *if* $S^t X \in C$, *for* $C \in \tilde{\mathcal{B}}$. *The point*

process N_C of times at which C-events occur is given by

$$N_C(B) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \in B, S^{T_n} X \in C), \qquad B \in \mathcal{B}$$

We are interested in probabilities of the form $P_t(S^t X \in G)$, where $G \subset C$ (for convenience). Sets of this type allow us to describe a wide variety of events associated with our Markov process. Learning how to deal with probabilities of this type allows us to study, for instance, the time that a Markov process spends in a subset, or the amount of time it takes a Markov process to travel from one subset to another.

To handle such a set C, we will need to assume that it is *decomposable*, in the sense that

$$\mathbf{1}(S^t X \in C) = \mathbf{1}(S^t X \in C^-_{X(t)}) \mathbf{1}(S^t X \in C^+_{X(t)}).$$

For each element $x \in \mathbb{E}$, C^-_x is a set in the σ-field generated by the collection of functions $\{\pi_t : t \in (\infty, 0)\}$, where $\pi_t : D(\Re) \to \mathbb{E}$ satisfies $\pi_t(z) = z(t)$ (i.e. it refers to the past with respect to time 0), and C^+_x is a set in the σ-field generated by $\{\pi_t : t \in [0, \infty)\}$. Notice also that $\mathbf{1}(S^t X \in C^-_x)$ is predictable with respect to the filtration generated by X. The following example will illustrate how this decomposition works.

Example 32 Suppose $C = \{z \in D(\Re) : (z(0-), z(0)) \in A\}$, where $A \subset \mathbb{E} \times \mathbb{E}$. Then for each $x \in \mathbb{E}$, $C^-_x = \{z \in D : (z(0-), x) \in A\}$. Notice that $\mathbf{1}(S^t X \in C^-_x)$ is predictable with respect to the natural filtration induced by X.

Theorem 33 *For $\mathcal{L}$-a.e. t,*

$$P_t(S^t X \in G) = \frac{\nu_G(t)}{\nu_C(t)},$$

where

$$\nu_C(t) = E\int_{\mathbb{E}} \mathbf{1}(S^t X \in C_x^-)P(X \in C_x^+ | X(0) = x)q(X(t), dx).$$

Proof Let N_C denote the point process of C-events (similarly for G-events). For any $B \in \mathcal{B}$, we see from the semi-regenerative property of X at the transition times, and the form of the stochastic intensity of M, that

$$\begin{aligned}
\mu_C(B) &= E\int_B \mathbf{1}(S^t X \in C)M(dt) \\
&= E\int_B \mathbf{1}(S^t X \in C^+_{X(t)}, S^t X \in C^-_{X(t)})M(dt) \\
&= E\sum_n P(T_n \in B, S^{T_n} X \in C^+_{X(t)}, S^{T_n} X \in C^-_{X(t)} | \mathcal{F}_{T_n}) \\
&= E\sum_n \mathbf{1}(S^{T_n} X \in C^-_{X(t)}, T_n \in B)P^{X(T_n)}(X \in C^+_{X(0)}) \\
&= E\int_B \mathbf{1}(S^t X \in C^-_{X(t)})P^{X(t)}(X \in C^+_{X(0)})M(dt) \\
&= E\int_B \int_{\mathbb{E}} \mathbf{1}(S^t X \in C_x^-)P(X \in C_x^+ | X(0) = x)q(X(t), dx)dt
\end{aligned}$$

This implies that μ_C is absolutely continuous with respect to $\mathcal{L}$, with density ν_C. Furthermore, notice that since $G \subset C$,

$$\begin{aligned}
\int_B P_t(S^t X \in G)\mu_C(dt) &= E\int_B \mathbf{1}(S^t X \in G)M(dt) \\
&= E\int_B \int_{\mathbb{E}} \mathbf{1}(S^t X \in G_x^-)P(X \in G_x^+ | X(0) = x)q(X(t), dx)dt
\end{aligned}$$

∎

In some cases, it is possible to simplify expressions of the form

$$E\int_B f(t, S^t X)N(dt)$$

without making explicit use of a stochastic intensity. Let

$$\xi(t) = \{X(s) : s < t\}$$

represent the "past" with regard to the time index t; similarly, let

$$\eta(t) = \{X(s) : s \geq t\}$$

represent the "present" and "future" with respect to t. We will also need to use another random variable $\tau_t = sup\{s < t : X(s) \neq X(t-)\}$. This variable represents the time of the last transition strictly before time t. Consider the functional

$$Z = \int_{\Re} f(t)g(\xi(\tau_t), \eta(t))M(dt) = \sum_n f(T_{n+1})g(\xi(T_n), \eta(T_{n+1}))$$

where $f : \Re \to \Re_+$, and $g : D(-\infty, 0] \times D((0, \infty)) \to \Re_+$. The use of f allows us to compute expectations for functionals of the form

$$Z_A = \int_A g(\xi(\tau_t), \eta(t))M(dt).$$

Expectations like these need to be computed for all Borel sets A in order to generate expressions for expectations like $E_t[g(\xi(\tau_t), \eta(t))]$.

By the strong Markov property, we can write

$$h(\xi(T_n), x) = E\left[g(\xi(T_n), \eta(T_{n+1}))|\mathcal{F}_{T_n}, X(T_{n+1}) = x\right],$$

where $h : D((-\infty, 0]) \times \mathbb{E} \to \Re_+$.

The following result is an extension of Lévy's formula. A similar result can be found in [29].

Theorem 34 *For the functional defined above,*

$$E[Z] = E\left[\int_{\Re}\int_{\mathbb{E}} h(\xi(\tau_t), x)q(X(t), dx)dt\right]. \tag{40}$$

Proof Conditioning on $\mathcal{F}_n = \mathcal{F}_{T_n}$ and $X_{n+1} = X(T_{n+1}) = x$, we have

$$E[Z] = \sum_n E[Z_n] \tag{41}$$

where

$$Z_n = \int_{\mathbb{E}} E\left[f(T_{n+1})g(\xi(T_n), \eta(T_{n+1}))|\mathcal{F}_n, X_{n+1} = x\right] P(X_{n+1} \in dx|\mathcal{F}_n).$$

Using 37 and 38, we see that

$$Z_n = \int_{\mathbb{E}} \left[h(\xi(T_n), x)E[f(T_{n+1})|\mathcal{F}_n, X_{n+1} = x]q(X_n, dx)/q(X_n)\right].$$

This follows from the fact that $T_{n+1} - T_n$ is conditionally independent of $\xi(T_n)$ and $\eta(T_{n+1})$, given $\mathcal{F}_n$ and $X_{n+1} = x$. Furthermore,

$$\begin{aligned}
E[f(T_{n+1})|\mathcal{F}_n, X_{n+1} = x] &= E\left[\int_0^\infty f(T_n + u)q(X_n)e^{-q(X_n)u}du|\mathcal{F}_n, X_{n+1} = x\right] \\
&= q(X_n)E\left[\int_0^\infty f(T_n + u)\int_u^\infty q(X_n)e^{-q(X_n)v}dvdu|\mathcal{F}_n, X_{n+1} = x\right] \\
&= q(X_n)E\left[\int_0^\infty \left[\int_0^v f(T_n + u)du\right] q(X_n)e^{-q(X_n)v}dv|\mathcal{F}_n, X_{n+1} = x\right] \\
&= q(X_n)E\left[\int_{(T_n, T_{n+1}]} f(u)du|\mathcal{F}_n, X_{n+1} = x\right].
\end{aligned}$$

Therefore,

$$E[Z_n] = E\left[\int_{(T_n, T_{n+1}]} f(u)du \int_{\mathbb{E}} h(\xi(T_n), x)q(X_n, dx)\right].$$

Substituting this in 41 yields 40. ∎

As we mentioned above, this result is an extension of a result known as Lévy's formula , which we now state as a corollary.

Corollary 35 *For $f : \Re \to \Re_+$ and $g : \mathbb{E}^2 \to \Re_+$,*

$$E\left[\int_{\Re} f(t)g(X(t-),X(t))M(dt)\right] = E\left[\int_{\Re}\int_{\mathbb{E}} f(t)g(X(t),x)q(X(t),dx)dt\right].$$

As a possible application, let's suppose we are interested in the amount of time a Markov process X spends in a finite set I. We will not assume that the process is in steady-state. Due to our interpretation of Palm probabilities, we see that the sojourn time probabilities of interest are $P_t(W(t) \leq x)$, where $W(t)$ is equal to the sojourn time spent in the set after the last transition into the set before t. Our intuition tells us that $W(t)$ should be phase-type under P_t, since we enter the set at precisely time t. Again, notice that

$$\begin{aligned} P_t(W(t) \leq x) &= \sum_{k\in I} P_t(W(t) \leq x | X_t = k) P_t(X_t = k) \\ &= \sum_{k\in I} F_k(x) P_t(X_t = k) \end{aligned}$$

where each F_k is the cumulative distribution function of a phase-type random variable. From [23], we see that $W(t)$ is phase-type under the measure P_t for $\mathcal{L}$-a.e. t, since it is just a finite mixture of phase-type distributions.

3.8 Little Laws

Consider a queueing system that operates as follows. Let N denote the point process on $\Re_+$ of times at which jobs arrive to the system. For now, assume that N is simple, and that there are no jobs in the system at time 0. Let $X(t)$ denote the number of jobs in the system at time t. The sojourn time in the system of job n will be denoted as W_n, and $W(s)$ will denote the waiting time in the system of the last job that arrived before or at time s.

For stationary systems, a classical Little law relates the expected queue length to the expected waiting time experienced by a customer. If our process (N, X) is stationary, and defined on $\Re$, we can say that

$$E[X(0)] = \lambda E^0[W(0)]$$

where λ is the rate of the stationary arrival process N.

Here we show that time-dependent expected queue lengths can also be written in terms of expected waiting times experienced by customers. This requires us to introduce a generalized notion of a Palm probability. Under our assumptions on $\mathcal{F}$, there exists a (μ_2-a.e. unique) probability kernel $P_{(t_1,t_2)}(A)$ such that

$$E[N(B)N(C)\mathbf{1}(\omega \in A)] = \int_B \int_C P_{(t_1,t_2)}(A)\mu_2(d(t_1,t_2))$$

where $\mu_2(B \times C) = E[N(B)N(C)]$. One can interpret $P_{(t_1,t_2)}(A)$ as the probability of A, given that N has points at t_1 and t_2. Furthermore, it's clear that this idea can be carried further to construct probabilities that condition on the locations of n points, for any integer $n \geq 1$.

Theorem 36 *(**Transient Little Law**) For any integer $n \geq 1$,*

$$E[X(t)^n] = \int_{(0,t]^n} P_{\boldsymbol{s}}(W(s_1) > t - s_1, W(s_2) > t - s_2, ..., W(s_n) > t - s_n)\mu_n(d\boldsymbol{s})$$

where $\mu_n(A_1 \times A_2 \times \cdots \times A_n) = E[\prod_{k=1}^n N(A_k)]$, for $A_k \in \mathcal{B}$, and $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)$.

Proof For clarity, assume that $n = 2$ (the other cases follow similarly). Notice that $X(t)$ can be written as follows:

$$X(t) = \sum_{n=1}^{\infty} \mathbf{1}(T_n \leq t, T_n + W_n > t) = \int_{(0,t]} \mathbf{1}(W(s) > t - s)N(ds).$$

Then, by repeated applications of the Campbell-Mecke formula and Lemma 11.2 in [16], we see that

$$E[X(t)^2] = \int_{(0,t]^2} P_{\boldsymbol{s}}(W(s_1) > t - s_1, W(s_2) > t - s_2)\mu_2(d\boldsymbol{s}).$$

∎

Now we will suppose that there are k customers waiting in the system at time 0. One can easily model this situation by letting N^* denote a point process with points $T_1 = T_2 = ... = T_k = 0$, and $0 < T_n < T_{n+1}$, for $n \geq k+1$ (N will consist of the points T_{k+1}, T_{k+2}, ...). Again, we do not assume that customers are served in any order (but we are still assuming customers arrive one at a time after time 0). In this case,

$$X(t) = \sum_{n=1}^{k} \mathbf{1}(W_n > t) + \int_{(0,t]} \mathbf{1}(W(s) > t - s)N(ds)$$

and so its second moment is as follows.

Corollary 37

$$\begin{aligned} E[X(t)^2] = & \sum_{n=1}^{k}\sum_{m=1}^{k} P(W_n > t, W_m > t) \\ & + 2\sum_{n=1}^{k} \int_{(0,t]} P_s(W_n > t, W(s) > t - s)\mu(ds) \\ & + \int_{(0,t]2} P_{\boldsymbol{s}}(W(s_1) > t - s_1, W(s_2) > t - s_2)\mu_2(d\boldsymbol{s}). \end{aligned}$$

Remark 38 Corollary 42 holds for any type of queueing system, regardless of the service discipline. Furthermore, if μ has a density h, then when $X(0) = 0$,

$$E[X(t)] = \int_0^t P_s(W(s) > t - s)h(s)ds.$$

Bertsimas and Mourtzinou [4] derived what is essentially this equation by using a sample-path approach. The first moment of $X(t)$ was derived using Palm theory in [24].

If we assume the service discipline of the queue is overtake-free, we can also derive a distributional form of Little's result using Palm measures.

Theorem 39 *(Distributional Little's Law) If the queueing system works under an overtake-free discipline, then*

$$P(X(t) \geq n) = \int_0^t P_s(W(s) > t - s, N(s,t] = n-1)\mu(dt).$$

Proof This follows by applying the Campbell-Mecke formula to

$$\mathbf{1}(X(t) \geq n) = \sum_{k=1}^{\infty} \mathbf{1}(W(T_k) > t - T_k, N(T_k,t] = n-1, T_k \leq t).$$

∎

Remark 40 In [4], this result was proved under the additional assumption that future arrivals do not influence the service times of all customers currently in the overtake-free system. In our Palm context, this means that, for μ-a.e.s, $W(s)$ is independent of N under P_s on the interval (s, ∞).

Corollary 41 *If, for μ-a.e.s, $W(s)$ is independent of N under P_s on the interval (s, ∞), then*

$$E[z^{X(t)}] = 1 + (z-1)\int_0^t P_s(W(s) > t - s)E_s[z^{N(s,t]}]\mu(ds).$$

Proof Under the independence assumption, we see from the result above that

$$P(X(t) \geq n) = \int_0^t P_s(W(s) > t - s) P_s(N(s, t] = n - 1) \mu(ds).$$

The generating function of $X(t)$ can now be easily derived with simple algebra.

∎

CHAPTER IV

APPROXIMATION OF JUMP PROCESSES

4.1 Introduction

Finding good approximations for queue-length processes is currently a very active area of research in applied probability. For example, it is known that under various heavy traffic regimes, queue-length processes often converge weakly to functions of a Brownian motion process. The closer the offered load is to one (hence the term heavy traffic), the better the approximation.

In our research, we are interested in approximating queue-length processes, without placing any assumptions on the parameters of the interarrival and service times associated with the system. In other words, we want to know what can be said about queues that do not fit within some sort of classical scaling regime. Our approximations will be with respect to how close queue-length processes are to each other, and this closeness needs to be quantified. Fortunately, our processes will take values in a special subset of $D(\Re_+)$, which consists of the set of right-continuous functions with left-hand limits.

Suppose $\{z_n\}_{n\geq 1}$, z are points in the space $D(\Re_+)$. There are many ways to measure how "close" z_n gets to z as $n \to \infty$. One of these ways involves using a metric that induces the Skorohod topology . References on this metric include [12, 34].

Definition 42 *We say that functions $z_n \to z$ in the Skorohod topology if, for each $T > 0$, there exists a sequence of strictly increasing continuous functions λ_n (possibly depending on T) such that $\lambda_n(0) = 0$, $\lim_{t\to\infty} \lambda_n(t) = \infty$, and*

$$\lim_{n\to\infty} \sup_{0\le t\le T} |z_n(\lambda_n(t)) - z(t)| = 0$$

$$\lim_{n\to\infty} \sup_{0\le t\le T} |\lambda_n(t) - t| = 0.$$

Let $D_0(\Re_+)$ denote the set of piecewise-constant functions that map from $\Re_+$ to $\mathbb{E}$ that only have a finite number of jumps in any compact set, where $\mathbb{E}$ is a metric space (equipped with the topology generated by the metric). Notice that if a sequence of functions $\{z_n\}_{n\ge1} \in D_0(\Re_+)$ is such that the corresponding discontinuity points $\{t_k^n\}_k$ converge to $\{t_k\}_k$ and the sequence $\{z_n(t_k^n)\}_k$ converges to $\{z(t_k)\}_k$, where $z \in D_0(\Re_+)$ is a function with discontinuities at $\{t_k\}_k$, then z_n converges to z in the Skorohod topology. This fact is a powerful one, as it allows us to simplify the problem of showing that functions converge to showing that vectors converge. More importantly, many of the well-known queueing systems have sample paths that lie in $D_0(\Re_+)$; this will prove to be very useful throughout the derivation of our results.

This simple observation leads to the following question: suppose Q^n is a sequence of queueing systems, where each system has associated with it a collection of primitives V^n (a collection of interarrival times, service times, routing variables, etc.). If V^n converges weakly to V, it is true that Q^n converges weakly to Q (with primitives V) with respect to the Skorohod topology? In general, the answer is no, as the next example illustrates.

Example 43 Consider a sequence of single-server queueing systems Q^n (that are empty at time 0) with deterministic interarrival times $U_1^n = 5$, $U_2^n = 4+1/n$, $U_k^n = 10$ for $n \geq 3$, and service times $S_1^n = 4$, $S_2^n = 1$, and $S_k^n = 7$ for $n \geq 3$. Clearly our primitives converge, but the corresponding queue-length processes cannot converge with respect to the Skorohod topology. This has to do with the fact that the limiting process doesn't have as many jumps as the original process, because an arrival and a service occur simultaneously at time 9.

This example suggests that if our queueing process is such that arrivals and departures never occur simultaneously, we can approximate it with a simpler queueing process under the Skorohod topology. However, if this assumption isn't satisfied, a weaker notion of convergence must be considered, if one hopes to arrive an an approximation result.

It should be mentioned that such a question has been studied before. For instance, Whitt [33] shows this result for queues that do not allow arrivals and departures to occur simultaneously. His method of proof makes heavy use of the fact that the queue-length at time t can be written as

$$Q(t) = A(t) - D(t).$$

where $A(t)$ and $D(t)$ denote the number of arrivals and departures in $(0, t]$, respectively.

Assuming that Q takes such a form, he then uses standard results from the theory of convergence of functions in $D[0, \infty)$ to arrive at the result. We will prove a more general statement, by taking advantage of the fact that $Q \in D_0(\Re_+)$. Our proof will

also give insight on how to handle the case when arrivals and departures are allowed to occur at the same time. This proof also shows, in a more elementary way, why we shouldn't hope for Skorohod convergence if we allow arrivals and departures to occur at the same time.

4.2 Jump Processes

Suppose $X = \{X(t)\}_{t \geq 0}$ is a stochastic process that takes values in a metric space $\mathbb{E}$ that's endowed with the topology of open sets. We'll say that X is a *Jump Process* if the sample paths of X reside in the space $D_0(\Re_+)$. We will denote the jump times of X as $\{T_k^{'}\}_{k \geq 0}$, with $T_0^{'} = 0$. Recall that membership in $D_0(\Re_+)$ implies that, with probability one, $\lim_{k \to \infty} T_k^{'} = \infty$. Notice that our index set is just $\Re_+$, and it will be very convenient to associate a notion of time to this set (as is traditionally done in the literature).

We will assume that X evolves through time as follows: at time $T_0 = 0$, the current state of X is just $X(0)$. Furthermore, at time 0, a collection of clocks $\mathcal{C}_0 \subset \mathcal{C}$ will begin operating, where $\mathcal{C}$ represents the set of all clocks, or the "clock space" of the process (we will assume that $\mathcal{C}$ is countable). We will assume that there exists a function $g : \mathbb{E} \to 2^{\mathcal{C}}$, where $g(X(t))$ represents the collection of clocks that are currently running at time t. At time 0, each clock $c \in \mathcal{C}_0$ will be assigned an initial value of $W_c(0)$, and the clock values will decrease through time at a rate of $r_c(X(0))$ until it reaches zero, at which time it "alarms".

These clocks will dictate when and how X transitions through the space $\mathbb{E}$. In particular, X will make a possible transition at an *action time* (i.e. a time at which

an alarm goes off). We denote these action points as the sequence $\{T_k\}_{k\geq 0}$, with $T_0 = 0$. For clarity of exposition, we will define these times in an inductive manner. The first action time is just the first time that an alarm is heard:

$$T_1 = \min_{c\in\mathcal{C}_0} \frac{W_c(0)}{r_c(X(0))}. \tag{42}$$

Let $\mathcal{C}_1^*$ denote the set of clocks that achieve the minimum in (42). Each clock $c \in \mathcal{C}_1^*$ triggers a new clock $R(c, \mathcal{C}_1^*, X(0))$ (which could be random) to begin operation with a time of $V_{R(c,\mathcal{C}_1^*,X(0))}(1)$, where $R(c, A, x)$ represents the clock that's triggered by an alarming clock c from the alarm set A, while the process is in state x, and $V_c(n)$ is the time allocated to clock c immediately after begin triggered for the n^{th} time. The set of clocks that begin operation at time T_1 is

$$\tilde{\mathcal{C}}_1 = \{R(c, \mathcal{C}_1^*, X(0)) : c \in \mathcal{C}_1^*\}.$$

Each of the clocks in $\mathcal{C}\backslash\mathcal{C}_1^*$ will assign its remaining time to another clock $h(c, \mathcal{C}_1^*, \tilde{\mathcal{C}}_1)$, where h is a function on $\{0, 1, 2, ...\} \times 2^{\mathcal{C}} \times 2^{\mathcal{C}}$ such that h is one-to-one from $\mathcal{C}\backslash\mathcal{C}_1^*$ to $\mathcal{C}\backslash\tilde{\mathcal{C}}_1$.

We will also require a set of clocks $\mathcal{C}_\pi$ to run continuously throughout the duration of the process (e.g., a clock representing arrivals into a queue). A clock in this set can trigger other clocks not in the set to begin (an arrival to an empty system typically triggers a service to begin), but clocks not in $\mathcal{C}_\pi$ cannot trigger clocks in $\mathcal{C}_\pi$. Accordingly, the assignment function satisfies $h(c) = c$ for $c \in \mathcal{C}_\pi$.

Then the set of clocks that are running at time T_1 is

$$\mathcal{C}_1 = \tilde{\mathcal{C}}_1 \cup \mathcal{C}_\pi \cup \{h(c, \mathcal{C}_1^*, \tilde{\mathcal{C}}_1) : c \in \mathcal{C}\backslash\mathcal{C}_1^*\}$$

and their wakeup values are

$$\begin{aligned} W_c(T_1) &= V_c(1), & c \in \tilde{\mathcal{C}}^1 \\ W_{h(c)}(T_1) &= W_c(0) - r_c(X(0))T_1, & c \in \mathcal{C}_0 \backslash \mathcal{C}_1^* \\ W_c(T_1) &= V_c(1), & c \in \mathcal{C}_\pi \cap \mathcal{C}_1^* \\ W_c(T_1) &= W_c(0) - r_c(X(0))T_1, & c \in \mathcal{C}_\pi \backslash \mathcal{C}_1^* \end{aligned}$$

and $W_c(T_1) = \infty$ for all other c's.

To complete our description of the event that occurs at time T_1, we will need a function f that dictates the transition behavior of X. In particular,

$$X(T_1) = f(X(0), \mathcal{C}_1^*, \tilde{\mathcal{C}}_1).$$

We will assume that $f(\cdot, A, B)$ is continuous, for any $A, B \in \mathcal{C}$.

As indicated above, the rest of the process can be described using induction. Suppose that we have constructed the process up to time T_n. Then the next action time T_{n+1} is just

$$T_{n+1} = T_n + \min_{c \in \mathcal{C}_n} \frac{W_c(T_n)}{r_c(X(T_n))}.$$

The set of clocks that alarm at time T_{n+1} is $\mathcal{C}_{n+1}^*$, and the set of clocks that begin operating at this time is just $\tilde{\mathcal{C}}_{n+1} = \{R(c, \mathcal{C}_{n+1}^*, X(T_n)) : c \in \mathcal{C}_{n+1}^*\}$. The next state visited is of course just $X(T_{n+1}) = f(X(T_n), \mathcal{C}_{n+1}^*, \tilde{\mathcal{C}}_{n+1})$.

In order to update the clock times at each action point, we will need to define a collection of point processes. Notice that the action points $\{T_k\}_{k \geq 1}$ induce a point process N, where

$$N(t) = \sum_{k \geq 1} \mathbf{1}(T_k \leq t).$$

Furthermore, for each $c \in \mathcal{C}$, two point processes will be defined:

$$\begin{aligned} N_c(t) &= \sum_{j=1}^{N(t)} \mathbf{1}(c \in \mathcal{C}_k^*) \\ \tilde{N}_c(t) &= \sum_{j=1}^{N(t)} \mathbf{1}(c \in \tilde{\mathcal{C}}_k). \end{aligned}$$

Notice that $N_c(t)$ represents the number of times clock c alarms in $(0, t]$, and $\tilde{N}_c(t)$ represents the number of times clock c is activated in $(0, t]$.

Thus, the clocks at time T_{n+1} are reset according to the following rule:

$$\begin{aligned} W_c(T_{n+1}) &= V_c(\tilde{N}_c(T_{n+1})), & c \in \tilde{\mathcal{C}}_{n+1} \\ W_{h(c)}(T_{n+1}) &= W_c(T_n) - (T_{n+1} - T_n) r_c(X(T_n)), & c \in \mathcal{C}_n \backslash \mathcal{C}_{n+1}^* \\ W_c(T_{n+1}) &= V_c(N_c(T_{n+1})), & c \in \mathcal{C}_\pi \cap \mathcal{C}_{n+1}^* \\ W_c(T_{n+1}) &= W_c(T_n) - (T_{n+1} - T_n) r_c(X(T_n)), & c \in \mathcal{C}_\pi \backslash \mathcal{C}_{n+1}^* \end{aligned}$$

and $W_c(T_{n+1}) = \infty$ for all other c's.

To summarize the preceding formulation, we say that the jump process X is represented by the *jump process system*

$$\zeta = (X, \{W_c\}, V, R, r, f, g, h, \mathcal{C}_\pi).$$

At this point, the reader may be confused as to how these action times relate to the jump times of X. For our applications, we will see that the set of jump times are a subset of the set of action times. In our queueing examples, it will be apparent that the action times may represent times at which an arrival and a service completion occur at exactly the same time, which would cause the queueing process to remain in the same state.

Below are some examples of queues that can be modelled as jump processes.

Example 44 ($FIFO$ Queues) A $FIFO$ queue can easily be modelled by the preceding framework. In this case, our clock set is just $\mathcal{C} = \{a, s\}$, where $R(a, \{a\}, 0) = s$, $R(a, \{a\}, 1) = 0$, $R(s, \{s\}, 1) = 0$, $R(a, \{a, s\}, 1) = 0$, $R(s, \{a, s\}, 1) = s$, and for $n \geq 2$, $R(a, \{a\}, n) = 0$, $R(s, \{s\}, n) = s$, $R(a, \{a, s\}, n) = 0$, and $R(s, \{a, s\}, n) = s$. We always want our arrival process to operate, so $\mathcal{C}_\pi = \{a\}$. Clearly $r_a(n) = \lambda$ for all $n \geq 0$, $r_s(0) = 0$, and $r_s(n) = \mu$ for all $n \geq 1$, and $f(0, \{a\}, \{s\}) = 1$, $f(1, \{a\}, \emptyset) = 2$, $f(1, \{s\}, \emptyset) = 0$, $f(1, \{a, s\}, \{s\}) = 1$ and for $n \geq 2$, $f(n, \{a\}, \emptyset) = n+1$, $f(n, \{s\}, \emptyset) = n - 1$, $f(n, \{a, s\}, \{s\}) = n$.

Example 45 ($LIFO$ queues) One can use exactly the same functions to model a nonpreemptive $LIFO$ queue as well.

Example 46 (Processor Sharing) Suppose we're interested in a $G/G/1$ queue operating under processor sharing. Under this queue discipline, all customers in the system are processed simultaneously, so we'll need more than two clocks. Our clock space can be chosen to be $\{a, s_1, s_2, s_3, ...\}$, and $\mathcal{C}_\pi = \{a\}$. The details of R are too cumbersome to list explicitly, but we will give a few examples to show how it's defined. For example, when there are three customers in the system, $\{a, s_1, s_2, s_3\}$ is the set of clocks that are running (to ensure this, we make sure that $\mathcal{C}_0$ is of this form). If a finishes first, then $R(a, \{a\}, 3) = s_4$. However, if $\{a, s_2\}$ alarm together, then $R(a, \{a\}, 3) = s_4$, $R(a, \{a, s_2\}, 3) = s_3$ (and $h(s_3) = s_2$), $R(s_2, \{a, s_2\}, 3) = 0$. In other words, the new arrival occupies clock 3, and all servers switch to lower clocks (so in this example we're actually using the h function). If for instance s_1 and s_2 are the clocks that achieve this minimum, then $R(s_1, \{s_1, s_2\}, 3) = R(s_2, \{s_1, s_2\}, 3) = 0$

and $h(s_3) = s_1$. The rate functions are obvious: $r_a(n) = \lambda$ for all $n \geq 0$, $r_{s_k}(n) = \mu/n$ for all $n \geq 1$, and $r_{s_k}(0) = 0$.

Notice that networks of such queues can be modelled using jump processes as well. In this case, our routing functions R are random. R characterizes the randomness present that controls how particles move from one queue to another. For instance, in a Jackson network, when a particle leaves a queue it chooses another queue based on a set of transition probabilities, and the choice of queue is independent of all other past information.

The idea of allowing a collection of clocks to govern how a process moves through time is not a revolutionary one. For instance, a very similar idea is used to described what is referred to as a generalized Semi-Markov process, and these are studied in the work of Schassberger [27, 28].

4.3 Continuity results

4.3.1 Skorohod convergence

We are now ready to state our main result. A similar result, within the context of the generalized Semi-Markov processes mentioned above, can be found in Whitt [34]. The statement of the theorem involves the notion of a sequence of functions $\{f^n\}_{n\geq 1}$ converging continuously to a function f, which we now define.

Definition 47 *We say that a sequence of functions* $\{f^n\}_{n\geq 1}$ converges continuously *to* f *if, for each sequence* $x_n \to x$ *as* $n \to \infty$, $f^n(x_n) \to f(x)$ *as* $n \to \infty$.

Theorem 48 *Let ζ, ζ_1, ζ_2, ... denote jump process systems as defined above. Suppose the primitives of ζ_n converge to those of ζ in that, as $n \to \infty$,*

$$(X_n(0), \{W_c^n(0)\}_c, V^n, \mathcal{C}_0^n, \mathcal{C}_\pi^n) \Rightarrow (X(0), \{W_c(0)\}_c, V, \mathcal{C}_0, \mathcal{C}_\pi) \tag{43}$$

and $(r^n, f^n, g^n, h^n, R^n)$ converges continuously to (r, f, g, h, R). Also, assume ζ is such that $|C_k^*| = 1$ for all $k \geq 1$ w.p.1. Then $X_n \Rightarrow X$ with respect to the Skorohod topology.

The proof of this result uses the following elementary fact.

Lemma 49 *Let $u, u_1, u_2, \ldots$ be vectors in $\Re^m$, and, for a fixed m, let*

$$I = \{i : u_i = \min_{1 \leq j \leq m} u_j\},$$

and define I^n similarly. If $u^n \to u$ as $n \to \infty$ and $|I| = 1$, then there exists an integer n_0 such that $I^n = I$, for $n \geq n_0$.

Proof Suppose $i \in I$, and let $\epsilon > 0$ be small enough so that $u_i < u_j - \epsilon$ for all $j \neq I$. Since u^n converges to u, there exists an integer n_0 such that $|u_j^n - u_j| < \epsilon/2$ for all $n \geq n_0$. Then for $j \neq I$ and all $n \geq n_0$,

$$u_i^n < u_i + \frac{\epsilon}{2} < u_j - \frac{\epsilon}{2} < u_j^n$$

which completes the proof. ∎

This lemma does not hold if the cardinality of I is larger than 1. For instance, take $u_k^n = 1/(n+m-k)$ for $1 \leq k \leq m$. It's clear that u^n converges to the zero vector,

but $I^n = \{1\}$ and $I = \{1, 2, ..., m\}$. This keeps us from using the Skorohod topology to approximate queues that allow arrivals and departures to occur simultaneously (recall the remark at the beginning of the section).

Proof It suffices to show by induction that, with probability one,

$$\lim_{n\to\infty}(T_k^n, X^n(T_k^n), W_c^n(T_k^n), C_k^n, N_c^n(T_k), \tilde{N}_c^n(T_k)) \tag{44}$$
$$= (T_k, X(T_k), W_c(T_k), C_k, N_c(T_k), \tilde{N}_c(T_k)), \qquad k \geq 0.$$

Clearly this is true for $k = 0$ by assumption (45). Now suppose it is true for some k. Then there exists an integer n_0 such that $C_k^n = C_k$ for all $n \geq n_0$. Thus, for such n, we see that

$$T_{k+1}^n = T_k^n + \min_{c\in C_k} \frac{W_c^n(T_k^n)}{r_c^n(X^n(T_k^n))}.$$

By Lemma 49, there exists an integer $n_1 \geq n_0$ such that for all $n \geq n_1$, C_{k+1}^{*n} is constant, so it converges to C_{k+1}^*. Since C_{k+1}^* is finite, there exists $n_2 \geq n_1$ such that $\tilde{C}_{k+1}^n$ is the same for all $n \geq n_2$, so this sequence also converges to $\tilde{C}_{k+1}$. Therefore,for all $n \geq n_1$, C_{k+1}^n converges to C_{k+1} (since the set C_π is the same for all n). This fact immediately shows that the sequences $N_c(T_{k+1}^n)$ converge to $N_c(T_{k+1})$ (and similarly for N_c^*). Our induction hypothesis also tells us that $W_c^n(T_{k+1}^n)$ also converges to $W_c(T_{k+1})$, and by continuity, we see that T_{k+1}^n converges to T_{k+1}. Finally, our continuity assumption of f allows us to conclude that $X^n(T_{k+1}^n)$ converges to $X(T_{k+1})$. Thus, (44) is true for $k + 1$, which completes the proof. ∎

4.3.2 A More General Convergence Result

As mentioned previously, to approximate queueing systems that allow arrivals and departures to occur simultaneously, we will need to consider a weaker notion of convergence.

We say that $z_n \to z$ in DJ (here DJ stands for Disappearing Jumps) if there exists a sequence of points $\{t_k^n\}_{k\geq 0}$ and a collection of sequences $\{\nu(k)^n\}_{k\geq 0}$ indexed by n such that the jump points of z_n are contained in $\{t_k^n\}_k$, and for each $k \geq 0$,

$$\lim_{n\to\infty} t^n_{\nu(k)^n} = t_k,$$

$$\lim_{n\to\infty} t^n_{\nu(k)^n+1} = t_{k+1},$$

and

$$\lim_{n\to\infty} z_n(t^n_{\nu(k)^n}) = z(t_k).$$

This notion of convergence allows us to consider a sequence of functions containing jumps that "disappear" in the limit. It is easy to show that the queue-length processes given in Example 43 converge in this sense.

Theorem 50 *Let ζ, ζ_1, ζ_2, ... denote jump process systems as defined above. Suppose the primitives of ζ_n converge to those of ζ in that, as $n \to \infty$,*

$$(X_n(0), \{W_c^n(0)\}_c, V^n, \mathcal{C}_0^n, \mathcal{C}_\pi^n) \Rightarrow (X(0), \{W_c(0)\}_c, V, \mathcal{C}_0, \mathcal{C}_\pi) \tag{45}$$

and $(r^n, f^n, g^n, h^n, R^n)$ converges continuously to (r, f, g, h, R). Also, assume that

$$\sup_{c,x} r_c(x) \leq M \tag{46}$$

Then there exists random times $\{\nu(k)^n\}_{k\geq 1}$ *such that, for each* $k \geq 0$,

$$T^n_{\nu(k)^n} \Rightarrow T_k,$$

$$T^n_{\nu(k)^n+1} \Rightarrow T_{k+1},$$

and

$$\begin{aligned} & (X_n(T^n_{\nu(k)^n}), \{W^n_c(T^n_{\nu(k)^n})\}_c, \mathcal{C}^n_{\nu(k)^n}, \{N^n_c(T^n_{\nu(k)^n})\}_c, \{N^n_c(T^n_{\nu(k)^n})\}_c) \\ \Rightarrow \quad & (X(T_k), \{W_c(T_k)\}_c, \mathcal{C}_k, \{N_c(T_k)\}_c, \{N^*_c(T_k)\}_c). \end{aligned}$$

Moreover, these random elements converge jointly in distribution.

We will require an extra condition on our process. Let $A : \overline{\Re}^{\mathcal{C}}_+ \to \overline{\Re}^{\mathcal{C}}_+$ be defined in the following way: for each $n \geq 0$,

$$A(\{\frac{W_c(T_k)}{r_c(T_k)}\}_c) = \{\frac{W_c(T_{k+1})}{r_c(X(T_{k+1}))}\}_c.$$

In other words, A just maps the clock values at time T_k to the clock values at time T_{k+1}. The reason we are introducing this notation is because we will require the queueing processes to satisfy a sort of preservation assumption. Consider, for example, the clock times $\{W_c(T_k)\}$ at time T_k. If $|\mathcal{C}^*_k| = m > 1$, we realize that m clocks will ring at time T_{k+1}. The preservation assumption says the following: suppose that a process is currently at time t_0 in state x with clock set C_x. If we know that a collection of clocks $\{c_1, c_2, ..., c_m\}$ (that could possibly all alarm simultaneously) have alarmed at times $t_{c_1}, t_{c_2}, ..., t_{c_m} > t_0$, and only those clocks have alarmed in the interval $(t_0, t]$, where $t = \max_{1\leq i\leq n} t_{c_i}$, then at time t, the process will be in state y with clock set C_y, and y is independent of t_{c_1}, ..., t_{c_m}.

To state this mathematically, the preservation assumption says that, for two given sequences $\{a_n\}_n$, $\{b_n\}_n$,

$$B_{A\{\frac{W_c(T_k)}{r_c(X(T_k))}\}_c}(A_n^m(\{\frac{W_c^n(T^n_{\nu(k)^n})}{r_c n(X_n(T_{\nu(k)^n}))}\}_c)) = A_n^m(\{\frac{W_c^n(T^n_{\nu(k)^n})}{r_c^n(X_n(T^n_{\nu(k)^n}))}\}_c)$$

where for two sequences $\{a_c\}_c$, $\{b_c\}_c$

$$B_{\{a_n\}_n}(\{b_n\}_n)_k = \begin{cases} b_k, & \text{if } a_k < \infty; \\ \\ \infty, & \text{if } a_k = \infty. \end{cases}$$

The reader should note that our jump process representations of $FIFO$, $LIFO$ and Processor Sharing queues satisfy this assumption.

Proof We will prove this result by induction on k. In particular, we will verify the theorem for the case when, for each n, $\nu(0)^n = 0$, and for each $k \geq 0$, $\nu(k+1)^n = \nu(k)^n + |\mathcal{C}^{*n}_{\nu(k)^n+1}|$. As before, assume that our initial conditions converge w.p.1. We assume that $\nu(0)^n = 0$ for all $n \geq 1$, and for $n \geq \hat{n}$, we find that

$$T_1^n = \min_{c \in C_0} \frac{W_c^n(0)}{r_c(X(0))}$$

converges to T_1 w.p.1, and so this proves the result for the case when $k = 0$. Now suppose that we know that for an integer k,

$$T^n_{\nu(k)^n} \to T_k, \qquad w.p.1$$

$$T^n_{\nu(k)^n+1} \to T_{k+1}, \qquad w.p.1$$

and

$$\begin{aligned} & (X_n(T^n_{\nu(k)^n}), \{W_c^n(T^n_{\nu(k)^n})\}_c, \mathcal{C}^n_{\nu(k)}, \{N_c^n(T^n_{\nu(k)^n})\}_c, \{N_c^n(T^n_{\nu(k)^n})\}_c) \\ \to \quad & (X(T_k), \{W_c(T_k)\}_c, \mathcal{C}_k, \{N_c(T_k)\}_c, \{N_c^*(T_k)\}_c) \qquad w.p.1. \end{aligned}$$

Due to the discrete nature of the clock sets, we know that there exists an n_0 such that $\mathcal{C}^n_{\nu(k)^n} = \mathcal{C}_k$, for all $n \geq n_0$. Let

$$\mathcal{C}^*_{k+1} = arg \min_{c \in \mathcal{C}_k} \left(\frac{W_c(T_k)}{r_c(X(T_k))} \right)$$

and let

$$\alpha = \min_{c \in \mathcal{C}_k} \left(\frac{W_c(T_k)}{r_c(X(T_k))} \right).$$

At each $c \in \mathcal{C}^*_{k+1}$, a new random variable $V^n_c(\tilde{N}^n_c(T^n_{\nu(k)^n} + 1))$ is selected as a new value for some alarm clock. Set

$$\delta = \min_{c \in \tilde{\mathcal{C}}_{k+1}} \frac{V_c(\tilde{N}_c(T_k) + 1)}{r_c(X(T_{k+1}))}.$$

and pick $n_1 \geq n_0$ such that for each $c \in \tilde{\mathcal{C}}_{k+1}$, and each $n \geq n_1$,

$$\frac{V^n_c(\tilde{N}^n_c(T^n_{\nu(k)^n}) + 1)}{M} \geq \frac{\delta}{2}$$

(recall (46)). Futhermore, there exists an $\eta \in (0, \delta/2)$ and an integer $n_2 \geq n_1$ such that, for each $c \in \mathcal{C}_k \backslash \mathcal{C}^*_{k+1}$,

$$\frac{W^n_c(T^n_{\nu(k)^n})}{r^n_c(X_n(T^n_{\nu(k)^n}))} \geq \alpha + \eta.$$

Futhermore, for any ϵ we choose that satisfies $0 < \epsilon < \eta$, there exists an integer $n_3 \geq n_2$ such that for each $c \in \mathcal{C}^*_{k+1}$,

$$\left| \frac{W^n_c(T^n_{\nu(k)^n})}{r^n_c(X_n(T^n_{\nu(k)^n}))} - \alpha \right| \leq \epsilon.$$

Therefore, for any $n \geq n_3$, the first $|\mathcal{C}^*_{k+1}|$ clocks that will ring after time $T^n_{\nu(k)^n}$ are the clocks that belong to the set $\mathcal{C}^*_{k+1}$ (a careful reader should realize that this statement

is very similar to what we needed to verify in the previous proof with the use of Lemma 49) , and $T^n_{\nu(k+1)^n+1} - T^n_{\nu(k+1)^n} \geq \eta + \alpha - (\alpha + \epsilon) = \eta - \epsilon > 0$. Thus, for such n we see that

$$T^n_{\nu(k+1)^n} = T^n_{\nu(k)^n} + \max_{c \in \mathcal{C}^*_{k+1}} \frac{W^n_c(X_n(T^n_{\nu(k)^n}))}{r^n_c(X_n(T^n_{\nu(k)^n}))} \to T_{k+1}$$

At this point, we can also conclude that for each c, both $N^n_c(T^n_{\nu(k+1)^n})$ and $\tilde{N}^n_c(T^n_{\nu(k+1)^n})$ converge to $N_c(T_{k+1})$ and $\tilde{N}_c(T_{k+1})$, respectively, since $N^n_c(T^n_{\nu(k+1)^n}) = N_c(T_{k+1})$ and $\tilde{N}^n_c(T^n_{\nu(k+1)^n}) = \tilde{N}_c(T_{k+1})$ for $n \geq n_3$. The preservation assumption also tells us that $X_n(T^n_{\nu(k+1)^n})$ converges to $X(T_{k+1})$, and that for each c, $W^n_c(T^n_{\nu(k+1)^n})$ converges to $W_c(T_{k+1})$.

Finally, it is easy to use what we have just proved to show that

$$T^n_{\nu(k+1)^n+1} = T_{\nu(k+1)^n} + \min_{c \in \mathcal{C}^n_{\nu(k+1)^n}} \frac{W^n_c(T^n_{\nu(k+1)^n})}{r^n_c(T^n_{\nu(k+1)^n})}$$

converges to T_{k+2}, and this completes the proof.

∎

4.4 Phase-type Approximation

Suppose that we are interested in the queue-length process $Q = \{Q(t) : t \geq 0\}$ corresponding to some queue of interest. Notice that if Q can be modelled as a jump process, then Theorems 48 and 50 imply that Q can be approximated in some sense by a sequence of queues Q^n, such that $(r^n, f^n, g^n, h^n, R^n) = (r, f, g, h, R)$, provided that both the primitives and the initial conditions of Q^n converge to those of Q.

Notice that in this case, the queue discipline of Q^n, for each $n \geq 1$ is the same as that of Q; only the primitives and initial conditions may be different.

A random variable ξ is *phase-type* if its distribution is that of an absorption time of a finite-state continuous time Markov chain. It is well known that, for any nonnegative random variable ξ, there exists a sequence of phase-type random variables ξ_n such that ξ_n converge weakly to ξ.

Our convergence results show that to approximate the queue-length process of a queue that possesses renewal interarrival and service times, we need only approximate the interarrival times and the service times with appropriate phase-type distributions. However, it should be mentioned that finding a good phase-type approximation to a random variable can truly be a difficult task, and there are many papers in the literature that are devoted to this topic. See, for example, [31]. Most of the literature is also devoted to fitting phase-type distributions to data. This makes our result very appealing, not only for the fact that it can be used to approximate any queue in theory, but also for the fact that many researchers are currently working on the problem of both efficiently collecting data on queueing systems, and fitting phase-type distributions to the data. For example, Brown et. al. [7] collect data on interarrival times, service times, waiting times, etc. from a small telephone call center.

Furthermore, it is also known (see [2]) that any point process on $\Re_+$ can be approximated by a sequence of Markovian arrival processes (while considering the weak topology on the space of measures). This allows us to approximate queues with dependencies among the interarrival times, and among the service times. The reader should notice that in our jump process models, we do not specify any sort of

independence structure amongst the primitives or the initial conditions.

4.5 Convergence to Markov jump processes

Even though we can now approximate any type of queue that fits within our jump process framework, there still may be problems with implementing such an approximation.

In order to approximate a random variable ξ with a phase-type random variable $\tilde{\xi}$, we need to know some more about the properties of ξ. At the very least, we should hope that there is data available that represent realizations of ξ. If so, we can use many of the known existing algorithms that create parameters for $\tilde{\xi}$ based on the data. Similarly, we would need this kind of information if we want to approximate a general queueing system with one that consists of primitives that are phase-type.

But what if such data does not exist, or is not available to us? Well, in this case we need some sort of universal conditions to exist such that our complicated queueing system is "close" in some sense to a much simpler queueing system when these conditions are satisfied. From the viewpoint of our continuity results, we should hope that if $\{Q_n\}_{n\geq 1}$ represents a sequence of queue-length processes converging to a simpler queueing process Q, our complicated process is one of the terms that are deep within the sequence. In other words, we'd like for our process to be equal in distribution to Q^{n_0}, for some very large n_0.

This sort of philosophy is very standard in the queueing approximation literature. Indeed, there are numerous papers in the literature that seek to approximate systems that satisfy certain conditions with simpler processes. These simpler processes are

typically functions of a Brownian motion, or some other type of diffusion.

In this section, we will attempt to find conditions under which a sequence of jump processes converges to a Markov jump process. To do this, we need to know precisely when a jump process is a Markov jump process. For instance, suppose that there exists a jump process such that $\{V_c(k)\}_{k\geq 1, c\in\mathcal{C}}$ and $\{W_c(0)\}_c$ are exponential random variables. Can we then conclude that the corresponding jump process X is a Markov jump process? It turns out that we can in this case, provided some extra conditions hold.

Theorem 51 *Suppose the jump process system ζ satisfies the following assumptions:*

(a) The random variables $V_c(n)$, for $c \in C$ and $n \geq 1$, are i.i.d. and are independent of $X(0)$ and $\{W_c(0)\}_c$.

(b) Conditioned on $X(0)$, the $W_c(0)$ is exponential with rate λ_c, and $V_c(n)$ is exponential with rate λ_c, for all $n \geq 1$.

(c) For any clock a, $\lambda_a = \lambda_{h(a)}$.

Then X is a Markov jump process with transition rate kernel

$$q(x, B) = \sum_{c\in g(x)} \lambda_c r_c(x) \mathbf{1}(f(x, c, R(c, \{c\}, x)) \in B), x \in \mathbb{E}, B \in \mathcal{E}.$$

Proof Suppose it were true that, on the set $\{X(T_k) = x\}$,

$$P(W_c(T_k) > w_c, c \in g(x)|\mathcal{F}_{T_k}) = e^{-\sum_{c\in g(x)} \lambda_c w_c}. \tag{47}$$

Then given $\mathcal{F}_{T_k}$, $W_c(T_k)$ is exponential with parameter λ_c, for $c \in g(x)$, so $W_c(T_k)/r_c(x)$ is exponential with parameter $\lambda_c r_c(x)$. Therefore, we see that on the set $\{X(T_k) = x\}$

$$P(X(T_{k+1}) \in B, T_{k+1} - T_k > t|F_{T_k})$$

$$
\begin{aligned}
&= \sum_{c \in g(x)} P(X(T_{k+1}) \in B, T_{k+1} - T_k > t, C^*_{k+1} = \{c\} | \mathcal{F}_{T_k}) \\
&= \sum_{c \in g(x)} P(f(x, c, R(c, \{c\}, x)) \in B, T_{k+1} - T_k > t, C^*_{k+1} = \{c\} | \mathcal{F}_{T_k}) \\
&= \sum_{c \in g(x)} \mathbf{1}(f(x, c, R(c, \{c\}, x)) \in B) \frac{\lambda_c r_c(x)}{q(x)} e^{-q(x)t}
\end{aligned}
$$

which proves that X is a Markov jump process.

It remains to prove (47). By property (b), we know that the result is true for $k = 0$. Now assume that it is true for some k. Our induction hypothesis tells us that, conditional on $\mathcal{F}_{T_k}$ and on the set $\{X(T_k) = x\}$, $W_c(T_k)$ is exponential, so there exists a unique clock c^* such that $C^*_{k+1} = c^*$. Let $\tilde{c} = R(c^*, \{c^*\}, x)$. Then on the set $\{X(T_k) = x\}$,

$$
\begin{aligned}
& P(W_c(T_{k+1}) > w_c, c \in g(X(T_{k+1})) | \mathcal{F}_{T_{k+1}}) \\
= \; & P(V_{\tilde{c}}(N_{\tilde{c}}(T_{k+1})) > w_{\tilde{c}}, W_c(T_{k+1}) > w_c, c \in g(X_{T_{k+1}}) - \tilde{c} | \mathcal{F}_{T_{k+1}}) \\
= \; & P(V_{\tilde{c}}(N_{\tilde{c}}(T_{k+1})) > w_{\tilde{c}}, W_{h(a)}(T_{k+1}) > w_{h(a)}, a \in g(X_{T_k}) - c^* | \mathcal{F}_{T_{k+1}}) \\
= \; & P(V_{\tilde{c}}(N_{\tilde{c}}(T_{k+1})) > w_{\tilde{c}}, W_a(T_k) - (T_{k+1} - T_k) r_a(X_{T_k}) > w_{h(a)} | \mathcal{F}_{T_{k+1}}) \\
= \; & e^{-\lambda_{\tilde{c}} w_{\tilde{c}}} e^{-\sum_{a \in g(X_{T_k}) - c^*} \lambda_a w_{h(a)}} \\
= \; & e^{-\lambda_{\tilde{c}} w_{\tilde{c}}} e^{-\sum_{a \in g(X_{T_k}) - c^*} \lambda_{h(a)} w_{h(a)}} \\
= \; & e^{-\lambda_{c \in g(x)} \lambda_c w_c}.
\end{aligned}
$$

Thus, (47) is true for $k + 1$, which completes the proof. ∎

With the preceding result in mind, we would like to find situations where our primitives converge to exponential random variables. Suppose we are interested in

scenarios where our sequence of interarrival times $\{U_k^n\}_k$ converges to an i.i.d. sequence of exponential random variables $\{U_k\}_k$. Let N_n denote a point process with points $T_k^n = \sum_{j=1}^k U_j^n$. Clearly we can think of N_n as a random element in the space $D(\Re_+)$. The following result relates convergence of N_n to convergence of $\{U_k^n\}_k$.

Proposition 52 *N_n converges weakly to N in $D(\Re_+)$ with respect to the Skorohod topology if and only if $\{T_k^n\}_k$ converges weakly to $\{T_k\}_k$.*

Proof Suppose N_n converges weakly to N. Then for any $m \geq 1$, and $t_1 < t_2 < ... < t_m$,

$$\begin{aligned}
\lim_{n\to\infty} P(T_1^n \leq t_1, T_2^n \leq t_2, ..., T_m^n \leq t_m) &= \lim_{n\to\infty} P(N_n(t_1) \geq 1, N_n(t_2) \leq 2, ..., N_n(t_m) \geq m) \\
&= P(N(t_1) \geq 1, N(t_2) \geq 2, ..., N(t_m) \geq m) \\
&= P(T_1 \leq t_1, T_2 \leq t_2, ..., T_m \leq t_m).
\end{aligned}$$

Thus, $\{T_k^n\}_k$ converges weakly to $\{T_k\}_k$.

Conversely, suppose the last statement holds. By the Skorohod representation theorem, we can assume the sequences converge w.p.1. Then N_n converges w.p.1. to N with respect to the Skorohod topology, since one can obviously pick increasing functions λ_n, where λ_n is defined by linear interpolation between points T_k and T_{k+1}, and $\lambda_n(T_k) = T_k^n$. ∎

Fortunately, there are results in the literature that provide conditions for a sequence of point processes $\{N_n\}_{n\geq 1}$ to converge to a Poisson process N. One of the usual assumptions is that each N_n can be written as the sum of point processes that

are "uniformly sparse". In other words, the point processes that make up each sum form what we will refer to as a null array, which we define below (see [15] for more details):

Definition 53 *Let* $\{N_{n,k}\}_{n\geq 1,k\geq 1}$ *be a collection of point processes on* $\Re_+$. *We say that this collection forms a* null array *if for each fixed* n, *the point processes* $\{N_{n,k}\}_k$ *are independent, and for each relatively compact set* B,

$$\lim_{n\to\infty} \sup_k E[|N_{n,k}(B)| \wedge 1] = 0.$$

Here is one such theorem (see [15]).

Theorem 54 *Let* $N_{n,k}$ *be a null array of point processes on* $\Re_+$, *and consider a Poisson process* N *on* $\Re_+$, *with intensity measure* μ. *Then*

$$\sum_k N_{n,k} \to N$$

if and only if (i) $\lim_{n\to\infty} \sum_k P(N_{n,k}(B) > 0) = \mu(B)$ *for all relatively compact sets* B *such that* $\mu(\partial B) = 0$.

(ii) $\lim_{n\to\infty} \sum_k P(N_{n,k}(B) > 1) = 0$ *for all relatively compact sets* B.

This result intuitively tells us that our continuity results should be useful when approximating queues who receive arrivals from many sources, each with long interarrival times. The reason we require each source to have long interarrival times is due to the fact that the point processes in the theorem form a null-array.

REFERENCES

[1] Asmussen, S. (2003). *Applied Probability and Queues*. Springer, New York.

[2] Asmussen, S. and Koole, G. (1993). Marked Point Processes as Limits of Markovian Arrival Streams. *J. of Appl. Prob.* **30**, 365-372.

[3] Baccelli, F. and Brémaud, P. (2003). *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurrences*. Springer, New York.

[4] Bertsimas, D. and Mourtzinou, G. (1997). Transient laws of non-stationary queueing systems and their applications. *Queueing Systems* **25** 115-155.

[5] Brémaud, P. (1989). Characteristics of queueing systems observed at events and the connection between stochastic intensity and Palm probability. *Queueing Systems* **5** 99-112.

[6] Brémaud, P. (1981). *Point Processes and Queues*. Springer, New York.

[7] Brown, L., Gans, N., Mandelbaum, A., Sakov, A, Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *JASA*. **100**, 36-50.

[8] Chung, K.L. (2001). *A Course in Probability Theory*. Academic Press, San Diego.

[9] Costa, O. L. V. and Dufour, F. (2005). A sufficient condition for the existence of an invariant probability measure for Markov Processes. *J. of Appl. Probab.* **42** 873-878.

[10] Dai, J. G. (1995) On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49-77.

[11] Daley, D. and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Springer, New York.

[12] Ethier, S. and Kurtz, T. (1986). *Markov Processes: Characterization and Convergence*. John Wiley and Sons, New York.

[13] Filonov, Y. P. (1990) Criterion for ergodicity of homogeneous discrete Markov chains. *Ukranian Math J.* **41** 1223-1225.

[14] Foss, S. and Konstantopoulos, T. (2004). An overview of some stochastic stability methods. *J. OR Soc. Japan* **47**, No. 4, 275-303

[15] Kallenberg, O. (2002). *Foundations of Modern Probability*. Springer, New York.

[16] Kallenberg, O. (1983). *Random Measures*. Akademie-Verlag, Berlin.

[17] Kleinrock, L. (1975). *Queueing Systems, Volume I: Theory*. John Wiley and Sons, New York.

[18] Last, G. and Brandt, A. (1995) *Marked Point Processes on the Real Line: The Dynamic Approach*. Springer, New York.

[19] Melamed, B. and Whitt, W. (1990). On arrivals that see time averages: a martingale approach. *J. of Appl. Probab.* **27** 376-384.

[20] Melamed, B. and Yao, D. (1995). The ASTA property in queueing. *Advances in Queueing: Theory, Methods, and Open Problems* (J. H. Dshalalow, Ed.) 195-224, CRC Press.

[21] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

[22] Meyn, S.P. and Tweedie, R. L. (1994). State-dependent criteria for convergence of Markov chains. *Ann. Appl. Probab.* **4** 149-168.

[23] Neuts, M. (1994) *Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Dover Publications, New York.

[24] Riano, G. (2002). *Transient Behavior of Stochastic Networks: Application to Production Planning with Load-Dependent Lead Times*. Ph. D. thesis, Georgia Institute of Technology.

[25] Robert, Philippe. (2003). *Stochastic Networks and Queues*. Springer-Verlag, Berlin.

[26] Rolski, T. (1989). Relationships Between Characteristics in Periodic Poisson Queues. *Queueing Systems* **4**, 17-26.

[27] Schassberger, R. (1978). Insensitivity of Steady-State Distributions of Generalized Semi-Markov Processes, Part I. *Ann. of Prob.* **5**, 87-99.

[28] Schassberger, R. (1979). Insensitivity of Steady-State Distributions of Generalized Semi-Markov Processes, Part II. *Ann. of Prob.* **6**, 85-93.

[29] Serfozo. R. (1999). *Introduction to Stochastic Networks*. Springer-Verlag, New York.

[30] Shiryayev, A. N. (1978). *Optimal Stopping Rules*. Springer-Verlag, New York.

[31] Thummler, A., Buchholz, P. and Telek, M. (2006). A Novel Approach for Phase-Type Fitting with the EM Algorithm. *IEEE Transactions on Dependable and Secure Computing*. **3**, 245-258.

[32] van Doorn, E.A., and Regterschot, G.J.K. (1988). Conditional $PASTA$. *Operations Research Letters* **7**, October No. 5, 229-232.

[33] Whitt, W. (1974). The Continuity of Queues. *Adv. Appl. Prob.* **6**, 175-183.

[34] Whitt, W. (1980). Continuity of Generalized Semi-Markov Processes. *Math. of Oper. Res.* **5**, 494-501.

[35] Whitt, W. (2002). *Stochastic-Process Limits.* Springer, New York.

[36] Wolff, R. (1982). Poisson arrivals see time averages. *Operations Research* **30** March-April No.2, 223-231.

INDEX